

MULTIDIMENSIONAL PRIVACY QUANTIFICATION FOR USER EMPOWERMENT

Dem Fachbereich Informatik der
Technischen Universität Darmstadt vorgelegte

Dissertation

zur Erlangung des akademischen Grades
Doktor-Ingenieur (Dr.-Ing.)
von

Spyridon Boukoros, Dipl.-Ing.,
geboren in Athen



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Referenten: Professor Dr. Stefan Katzenbeisser
Professor Dr. Carmela Troncoso
Professor Dr. Max Mühlhäuser

Tag der Einreichung: 25.02.2019

Tag der Prüfung: 10.04.2019

Darmstadt, 2019

Spyridon Boukoros: *Multidimensional privacy quantification for user empowerment* , February 2019

This work is licensed under a [Creative Commons "Attribution-NonCommercial-NoDerivatives 4.0 International"](#) license.



"Our life is what our thoughts make it."

— Marcus Aurelius, *Meditations*

ABSTRACT

As we are living in an interconnected world, serious privacy concerns have been raised due to the ever increasing data collection. As a result, many privacy-preserving methodologies have been proposed for various domains. However, in order to argue about the effectiveness of such methodologies, it is necessary to quantify privacy. Furthermore, such a quantification should be within the grasp of users and developers, as an important factor of any technology is adoption. In this thesis, we design privacy metrics in order to investigate the privacy guarantees offered by various defenses. In addition, we develop tools that can be used either by system designers for developing more privacy-preserving applications, or by users to estimate their privacy. The analysis is performed on three domains where millions of users already contribute data.

The first part of this thesis investigates a previously unexplored dimension of location privacy: mobile crowdsourcing, where users share streams of their location data. In this thesis we shed light as to whether traditional location privacy mechanisms can be directly applied in this scenario. We elaborate on why this use case is radically different than the widely studied case of location-based services. Then, using novel privacy metrics, and realistic utility functions and datasets, derived directly from crowdsourcing projects, we highlight why existing privacy defenses are inadequate. In order to enable further research in this direction and spawn privacy-preserving crowdsourcing applications, we provide some best-practices guidelines, directions for the development of novel defense mechanisms, and we show how our work can be used as a tool to measure privacy and utility loss.

In the second part of the thesis, we explore the privacy guarantees of aggregation schemes in smart metering, by modeling privacy as an indistinguishability game. In particular, we explore how many household have to be aggregated in order to provide meaningful privacy guarantees. We explain why such a modeling is flexible, able to simulate a variety of adversaries with different background information, and how the proposed game can be re-purposed to investigate as to whether single profiles belong in an aggregate or not. We investigate various aggregation sizes for privacy leakage, as well as, properties of an aggregate that affect the privacy guarantees.

The last part of this thesis investigates privacy in microdata publication. A tool is proposed, that enables users to estimate their privacy level, based on a set of preferences they want to share with a service provider (eg., movies watched, music listened to etc.), a-priori sharing them. The tool does not require full access to the providers' database

but rather relies on users' choices and the popularity of those. We describe the underlying privacy metric and the algorithms composing the tool. Using actual user data and comparing the tool's results with a well established privacy metric, we show that the tool is able to approximate users' privacy levels.

The privacy evaluation in the domain of mobile crowdsourcing highlights that the utility functions of the domain are different than those used in traditional location-privacy literature. For this reason, the utility-privacy trade-off of various defenses is different than the one observed in the scenario of location-based services, allowing us to understand why existing defenses are not deployed in practice by crowdsourcing projects. For the domain of smart metering, our work illustrates that aggregation-based privacy mechanisms are inadequate for small or medium sized aggregates of electricity consumption data. Users' electricity consumption patterns can be quite distinct, and with some auxiliary information sensitive data can be leaked from the aggregated report. Last, the user-friendly tool proposed for the domain of microdata publications, as well as the metrics and tools developed for the domains of mobile crowdsourcing and smart metering, can enable non-technical users to better understand the privacy risks of sharing unprotected data, and guide application developers towards developing privacy-preserving systems.

ZUSAMMENFASSUNG

Da wir in einer vernetzten Welt leben, treten vermehrt ernsthafte Bedenken hinsichtlich des Datenschutzes aufgrund der ständig zunehmenden Datenerhebung auf. Infolgedessen wurden viele Methoden zum Schutz der Privatsphäre für verschiedene Domänen vorgeschlagen. Um jedoch die Wirksamkeit dieser Methoden zu zeigen, ist es erforderlich die Privatsphäre zu quantifizieren. Darüber hinaus sollte eine solche Quantifizierung unter Kontrolle der Nutzer und Entwickler liegen, da Verwendbarkeit ein wichtiger Faktor für jede Technologie ist. In dieser Arbeit entwerfen wir Datenschutzkennzahlen, um die Datenschutzgarantien zu untersuchen, die verschiedene Schutzmechanismen bieten. Darüber hinaus entwickeln wir Werkzeuge, die entweder von Systementwicklern für die Entwicklung weiterer Datenschutz-Tools verwendet werden können oder von Benutzern, um ihre Privatsphäre einzuschätzen. Die Analyse wurde in drei Datensammlungen durchgeführt, zu denen Millionen von Benutzern bereits Daten beitragen.

Der erste Teil dieser Arbeit untersucht eine bisher unerforschte Dimension des Datenschutzes von Standorten: mobiles Crowdsourcing, bei dem Benutzer Streams ihrer Standortdaten gemeinsam nutzen. In dieser Arbeit untersuchen wir, ob traditionelle Mechanismen für den Schutz von Standortdaten direkt auf dieses Szenario angewendet werden können. Wir erläutern, warum sich dieser Anwendungsfall grundlegend von dem weithin untersuchten Fall ortsbasierter Dienste unterscheidet. Dann verwenden wir neuartige Datenschutzkennzahlen, realistische Hilfsfunktionen und Daten, die direkt aus Crowdsourcing-Projekten stammen, um zu zeigen, dass heutige Verteidigungsmechanismen der Privatsphäre unzureichend sind. Um weitere Forschung zu animieren Privatsphäre schützende Crowdsourcing-Anwendungen hervorzubringen, stellen wir einige Best Practice-Richtlinien sowie Anweisungen für die Entwicklung neuartiger Abwehrmechanismen vor, und wir zeigen, wie unsere Arbeit als Instrument zur Messung der Privatsphäre und des einhergehenden Verlustes von Nutzen verwendet werden kann.

Im zweiten Teil der Arbeit untersuchen wir die Garantien für den Datenschutz von Aggregationsverfahren in Smart Metering durch Modellierung der Privatsphäre als Ununterscheidbarkeits-Spiel. Insbesondere untersuchen wir, wie viele Haushalte zusammengefasst werden müssen, um angemessene Privatsphäre zu gewährleisten. Wir erklären, warum eine solche Modellierung flexibel ist sowie eine Vielzahl von Angreifern mit unterschiedlichen Hintergrundinformationen simulieren kann und wie das vorgeschlagene Spiel wieder ver-

wendet werden kann, um zu prüfen, ob ein Profil zu einem Aggregat gehört oder nicht. Wir untersuchen verschiedene Aggregationsgrößen auf das Eindringen in die Privatsphäre sowie welche Eigenschaften eines Aggregats die Datenschutzgarantien beeinflussen.

Der letzte Teil dieser Arbeit untersucht den Datenschutz bei der Veröffentlichung von Mikrodaten. Es wird ein Tool vorgeschlagen, mit dem Benutzer das Privatsphäre-Niveau ihrer Sammlung von Präferenzen (z. B. angesehene Filme, gehörte Musik, usw.) einschätzen können, um zu beurteilen, ob sie diese mit den Diensteanbietern teilen möchten. Das Tool erfordert jedoch keinen vollständigen Zugriff auf die Datenbank der Anbieter sondern verlässt sich auf die Auswahl der Benutzer und deren Beliebtheit. Wir beschreiben die zugrunde liegende Datenschutzmetrik und die Algorithmen, aus denen das Tool besteht. Anhand von tatsächlichen Benutzerdaten und dem Vergleich der Ergebnisse des Tools mit einer bereits etablierten Datenschutzkennzahl zeigen wir, dass das Tool eine Schätzung des Datenschutz-Niveaus ermöglicht.

Die Bewertung der Privatsphäre im Bereich des mobilen Crowdsourcing hebt hervor, dass die Utility-Funktionen der Domäne anders sind als diejenigen, die in der traditionellen Literatur zum Thema Standortdatenschutz verwendet werden. Aus diesem Grund unterscheidet sich der Kompromiss von Nutzen und Privatsphäre zwischen den verschiedenen Schutzmechanismen von dem, der im Szenario ortsbasierter Dienste beobachtet wurde. Das erlaubt uns zu verstehen, warum bestehende Abwehrmechanismen in der Praxis nicht in Crowdsourcing-Projekten eingesetzt werden. Für den Bereich Smart Metering zeigt unsere Arbeit, dass aggregationsbasierte Mechanismen zum Schutz der Privatsphäre für kleine oder mittlere Aggregate von Stromverbrauchsdaten unzureichend sind. Die Stromverbrauchsmuster der Benutzer können sehr unterschiedlich sein, sodass mit einigen Hilfsinformationen vertrauliche Daten aus dem Aggregat abgeleitet werden können. Das vorgeschlagene benutzerfreundliche Werkzeug für die Domäne von Mikrodatenpublikationen sowie die entwickelten Metriken und Werkzeuge für die Bereiche mobiles Crowdsourcing und Smart Metering können nicht-technischen Benutzern ein besseres Verständnis der Datenschutzrisiken bei ungeschützter Freigabe von Daten ermöglichen und leiten Anwendungsentwickler in Richtung Entwicklung von Systemen zum Schutz der Privatsphäre.

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my advisor Prof. Stefan Katzenbeisser for giving me the opportunity to join his group and guiding my research. It is common knowledge that a Ph.D. journey can be tough sometimes, with lots of ups and downs. Luckily, he was always there to make things easier. I would also like to thank my colleagues in the SecEng group for all the fun we had the last years. Especially, I would like to thank Niklas Büscher for helping me on numerous occasions during my first years in research. I would also like to thank Sheikh Mahbub Habib and Jörg Daubert for the many helpful research discussions we had.

During my Ph.D., I visited EPFL for three months and worked under the guidance of Prof. Carmela Troncoso. I am grateful to her for introducing new exciting research topics to me, and a great collaboration. Working closely with her has benefited me both as a researcher and as a person. Furthermore, I would like to thank EPFL's SPRING group for their hospitality during those three months. It was a memorable research and personal experience.

During my conference trips and my research visit to EPFL, I met and collaborated with Igor Bilogrevic and Mathias Humbert. I cannot thank them enough for all the research discussions we had that helped shaping this thesis and other publications.

The last years I have been lucky enough to meet many incredible people and share a large part of my daily life with them. I would like to thank Carel for having an amazing time both in and out of office. Also, thanks for teaching me so many stuff and mentoring me in various aspects of my life. During a rough period, I was lucky to share an office with Salman. That alone made things seem much easier. Many thanks also to Nikos, Pavlos, George and Stelios for all the fun times! I hope we will have many more.

Special thanks to my parents Panagiotis and Sofia, and my sister Vicky. Their unconditional love and support always leaves me speechless. Last but by no means least, I would like to thank Christina. This thesis would not have been possible without her support and love. Thanks for being there on both good and bad days.

CONTENTS

1	INTRODUCTION	1
2	PRIVACY IN MOBILE CROWDSOURCING APPLICATIONS	9
2.1	Mobile crowdsourcing	9
2.2	Related work	11
2.2.1	POIs and the uniqueness of human mobility . .	11
2.2.2	Privacy mechanisms	12
2.2.3	Attacks on geo-located data	14
2.2.4	Privacy quantification	15
2.3	MCS applications under study	15
2.3.1	Safecast	16
2.3.2	Radiocells	17
2.4	Privacy risks in current MCS applications	18
2.4.1	Safecast	19
2.4.2	OpenStreetMaps	23
2.5	Protecting location privacy in MCS	23
2.5.1	Defenses	24
2.5.2	Measuring privacy	25
2.5.3	Measuring utility	26
2.6	Existing LPPMs performance in MCS scenarios	27
2.6.1	Experimental setup	27
2.6.2	Defenses implementation	28
2.6.3	Privacy gain	31
2.6.4	Utility-privacy trade-off	37
2.7	Towards privacy-preserving mobile crowdsourcing . .	46
2.7.1	A tool for the systematic evaluation of LPPMs for MCS	47
2.7.2	Towards effective defenses	48
2.7.3	Best practices for mobile crowdsourcing developers	48
2.8	Chapter Summary	49
3	PRIVACY ASSESSMENT OF AGGREGATION SCHEMES IN SMART METERING	51
3.1	Smart metering	51
3.2	Related work	53
3.2.1	Privacy mechanisms for smart metering	53
3.2.2	Privacy quantification	55
3.3	Aggregation privacy model	56
3.3.1	Smart grid aggregation model	56
3.3.2	Requirements of privacy notions for aggregation in the smart grid	56
3.3.3	Smart grid privacy model	57
3.4	Methodology	60
3.4.1	Smart grid datasets	61

3.4.2	Evaluation approach	62
3.4.3	Decision functions	63
3.5	Case studies	64
3.6	Dataset analysis	75
3.7	Chapter Summary	79
4	MEASURING PRIVACY IN HIGH DIMENSIONAL MICRO-DATA COLLECTIONS	81
4.1	Microdata releases and privacy	81
4.2	Related work	83
4.2.1	Reidentification attacks	83
4.2.2	Privacy quantification	84
4.3	Privacy assessment mechanism	85
4.3.1	Overview of the system	85
4.3.2	System's internal logic	85
4.4	Evaluation	90
4.4.1	Datasets	90
4.4.2	Designing the experiments	91
4.4.3	Validity of the model	94
4.5	Privacy thermometer	97
4.6	Chapter Summary	98
5	CONCLUSION	99
	BIBLIOGRAPHY	103

LIST OF FIGURES

Figure 2.1	Cluster size and amount of POIs per user. . . .	22
Figure 2.2	GeoInd noise magnitude.	29
Figure 2.3	Safecast privacy gain.	30
Figure 2.4	Precision and recall results in Tokyo.	33
Figure 2.5	Clusters' size and points of interests (POIs) per cluster in Tokyo.	34
Figure 2.6	Spatial precision and recall for Safecast.	35
Figure 2.7	POIs precision and recall for Safecast.	35
Figure 2.8	Radiocells privacy gain.	36
Figure 2.9	Spatial precision and recall for Radiocells. . . .	37
Figure 2.10	POIs precision and recall for Radiocells.	37
Figure 2.11	Measurement error in Tokyo using a distance-based metric.	38
Figure 2.12	Measurement error in Tokyo's radiation values. .	39
Figure 2.13	Measurement error in Fukushima's radiation values.	39
Figure 2.14	Users' prior probabilities in Tokyo.	41
Figure 2.15	Hotspot detection in Tokyo.	44
Figure 2.16	Hotspot detection in Fukushima.	45
Figure 2.17	Utility loss in Radiocells.	46
Figure 2.18	Location privacy framework.	47
Figure 3.1	The proposed privacy game AggG.	58
Figure 3.2	Smart metering load profiles.	65
Figure 3.3	Comparison of the four decision functions. . .	66
Figure 3.4	The effect of the window parameter on a decision function.	67
Figure 3.5	Distinguishing advantage for the UMASS, GO-VAU, and Dataport datasets.	67
Figure 3.6	Comparison of the decision functions for different aggregation sizes.	68
Figure 3.7	Impact of temporal resolution.	69
Figure 3.8	The combined decision function over different aggregation sizes.	70
Figure 3.9	Imprecision for too large aggregation sizes. . .	71
Figure 3.10	Detectability of single appliances.	72
Figure 3.11	Distinguishability of various appliances.	73
Figure 3.12	Correlation of device properties and adversarial advantage.	74
Figure 3.13	Indistinguishability game vs membership inference.	75
Figure 3.14	Mean load profiles for the datasets.	76

Figure 3.15	Distance between the mean and every load profile.	77
Figure 3.16	Correlation of profiles with the average.	77
Figure 3.17	Mean load profile for the generated dataset. . .	79
Figure 3.18	Comparison of the decision functions for the generated dataset.	80
Figure 4.1	The proposed privacy quantification tool. . . .	86
Figure 4.2	Number of items in user profiles.	92
Figure 4.3	Items' popularity.	93
Figure 4.4	Evaluation of the tool on the Yahoo! Movies dataset.	94
Figure 4.5	Evaluation of the tool on the Yahoo! Music dataset.	95
Figure 4.6	Evaluation of the tool on the Netflix dataset. .	95
Figure 4.7	EU energy efficiency label.	97
Figure 4.8	Example of privacy label for the Netflix dataset.	98

LIST OF TABLES

Table 2.1	Safecast API dataset statistics.	19
Table 2.2	Safecast and Radiocells measurements per region.	27
Table 2.3	Category changes, for each of the radiation danger groups, after applying Geo-Ind 300 in Fukushima.	42
Table 2.4	Category changes, for each of the radiation danger groups, after applying the Rounding 2 in Fukushima.	42
Table 3.1	Smart grid datasets.	61
Table 3.2	Usable buildings per dataset.	62
Table 3.3	Adversarial advantage for different timeslots. .	70
Table 3.4	Characteristic properties of particular appliances (average values).	72
Table 4.1	Microdata datasets.	91
Table 4.2	Comparison of the average anonymity set and user's privacy levels.	96

ACRONYMS

MSE mean squared error

MLE	maximum likelihood estimation
NILM	non-intrusive appliance load monitoring
POI	points of interest
LBS	location-based service
MCS	mobile crowdsourcing
GeoInd	geo-indistinguishability
LPPM	location privacy-preserving mechanism
GPS	global positioning system

INTRODUCTION

The need for privacy has always been relevant. From the ancient Greek times there has been a distinction between the public and the private. However, advances in technology played an important role to the shape of later privacy definitions. A very early definition of privacy was by Judge Cooley in 1879, referring to it as the right to be left alone [51]. In a later definition in 1967, Westin refers to privacy as individual's ability to control or delete information about themselves as well as, decide how such information is communicated [154]. The advances in telecommunications though, have changed the way information is transmitted, stored and processed. Nowadays, Nissenbaum's work on contextual integrity, amongst others, frames privacy as an appropriate flow of information and recognizes five key parameters: data subject, sender, recipient, information type, and transmission principle [117].

Today's massive and detailed collection of information about individuals violates their privacy and in order to preserve it, only relevant and the minimum information necessary for specific tasks needs to be collected. The collected information though, should not allow the inference of unintended sensitive information. However, preventing inference attacks from published data turns out to be a rather complicated issue, and the main motivating problem of this thesis.

Humans are social beings and belong to many different groups and social circles and hence, it is only natural that with the widespread adoption of the Internet, such behavior would extend to the web. Indeed, depending on their interests, people participate in various online communities, forums and web-platforms. Contrary to real life though, where people (usually) carefully choose what information they share with whom, such an attitude is not yet common on the web. In addition, it is not usually clear to people what kind of sensitive information published data carry.

It is known that as little as one's queries to online search engines reveal her interests [131, 138, 147]. Lately, people have been outsourcing their preferences, photos, videos or even DNA test results to social media. Such data however can be used for discrimination. It is known that recruiters and hiring managers systematically search applicants social media profiles [43, 118]. More specifically, candidates have been disqualified because of posts they publish on their profiles, unprofessional screen names, inappropriate photos or too frequent posting [72]. Furthermore, advances in machine learning and the abundance of available data enable rather dangerous inferences. It has been demon-

strated that sensitive data such as religion, sexual preference, location or health history can be inferred directly or indirectly from someone's online activity [96]. Moreover, the photos someone publishes can be used to predict markers of depression [133].

Private personal information can also be inferred from shared location data. Smartphones have embedded global positioning system (GPS) receivers or use the device's sensors for localization purposes. An increasing number of applications use the device's position to offer location-based services (LBSs). Such services may send notifications about nearby friends, interesting locations based on users' preferences and weather or traffic alerts. Furthermore, besides assisting users, various companies collect useful statistics from users' whereabouts. Such measurements, however, carry extremely sensitive data to providers, as workplaces or home addresses can be determined [74, 97]. Moreover, social relationships as well as routines can also be inferred [36, 75]. For example, being in close proximity with someone during weekend nights might suggest that this person is a friend, while a romantic involvement can be inferred (with some probability) by people spending several nights together. Even though various anonymization techniques might be applied to location datasets, re-identification attacks are still possible because, as little as the combination of home and work location, or the shops one visited, is enough to uniquely identify individuals [56, 57, 157].

Smartphones however, are not the only devices with Internet connection capabilities and sensors. Nowadays, many other devices, capable of measurements and an Internet connection, are part of modern houses. Smart devices that measure the electricity consumption in households can pose a serious privacy threat. Such devices collect data regarding the electricity usage of all appliances in a household. Even though such a data collection has not a malicious purpose, but rather to allow energy providers to better manage the distribution of energy, post-processing of the collected data can reveal sensitive information about the owner. For instance, by looking into the usage patterns of such smart devices, one can infer when a building is occupied, how many people are present, peoples' sleeping patterns or even movie and music preferences.

As it is clear that individuals' privacy can be violated by various technologies used in everyday life, many privacy defenses have been proposed for a variety of domains. Such privacy-preserving technologies frequently rely (but not only) on cryptography, aggregation, selective hiding, anonymization, noise addition or generalization. However, in order to argue about the effectiveness of privacy preserving technologies, there is a need to define what privacy is and ways to quantify it. Even though various privacy definitions are abstract enough to define privacy in many domains, such as indistinguishability or adversarial error, the specifics of each area, such as the nature

of the collected data, are quite distinct. For instance, measuring how private one is when sharing his location data, which is in the form of coupled coordinates with time, is different than measuring one's privacy in microdata publication, which is in essence a collection of his preferences in a domain. Furthermore, as it will become clear in Chapter 2, even collected data of the same nature require different treatment depending on the resolution in which they were gathered. Hence, it follows naturally that privacy quantification must be treated as a domain dependent problem.

However, simply proposing privacy metrics and tools does not guarantee that they will be used. A fundamental problem of any proposed technology is adoption. First, researchers need to demonstrate the privacy dangers involved in the publication of raw user data in order to sensitize users, and deploy ready to use privacy-enhancing technologies. In order to sensitize users, the proposed privacy tools need not only to inform the users about their privacy levels, but through them raise awareness, something that can happen by using understandable concepts to measure privacy and usability friendly techniques such as visual results. Thus, once users are educated about privacy and start demanding more privacy-preserving tools in their daily life, privacy can be incorporated in industrial production pipelines. Hence, application developers and service providers should also be able to grasp privacy quantification results and have access to intuitive tools. Thus, there is also a need for privacy measurement frameworks that can easily be extended by them according to their needs.

CONTRIBUTION

This thesis focuses on privacy quantification. As already discussed, privacy is domain dependent and hence, we focus on three domains selected based on the amount of current or potential data contributors. Namely, these domains are mobile crowdsourcing (MCS), smart metering, and microdata collection. The domains of MCS and microdata datasets have millions of data contributors as location and preferences collection through mobile devices and online services respectively, are widely deployed. Smart metering is still at an early stage with experimental projects running on various areas around the world, however, it might soon be the standard way of electricity measuring for millions of users. For each of the aforementioned domains, we collected publicly available data from actual service providers. We designed privacy metrics and developed tools based on easily understandable concepts by users and practitioners. Then, we evaluated users' privacy by performing attacks on the collected data. Finally, as a very important factor of the successful application of any privacy-protection mechanism is the utility degradation of the data protected,

we evaluated the privacy-utility trade-off whenever we had access to the datasets' utility functions.

Our contributions are as follows:

1. We perform the first in-depth study on location privacy in the context of MCS. Contrary to the existing paradigm of LBSs, where users reveal their location seldom to the provider, in the case under study users upload continuously measurements. In order to understand the extent of the privacy danger, we perform various inference attacks (e.g., identifying workplaces, social relationships, and even sensitive information such as religious views) on the collected data from two representative applications which make their contributors' data publicly available on their websites. Since MCS is a previously unexplored area of location privacy, we investigate if and to which extend previously proposed location privacy-preserving mechanisms (LPPMs) can be used. To this end, we propose two privacy metrics, based on statistical measures developed for binary classification and information retrieval, that capture the privacy gain provided by the LPPMs with respect to the identification of areas and points of interest. Furthermore, we also consider new utility measures that quantify the accuracy of the aggregated values of data collectively gathered at these locations. Our experimental evaluation on real data shows that mechanisms based on sharing only a part of users' location history do not bring any privacy benefits to MCS users, essentially because the geo-located data is in general reported over a long period of time (more than a day) and contains too many locations. Mechanisms that change the reported position to a random one close to the original, tend to provide better privacy but only for strong privacy parameters that in turn significantly affect the resulting utility. Techniques that reduce the localization precision though, seem to be the best suited mechanisms privacy-wise. However, we show that none of the LPPMs provide an acceptable privacy-utility trade-off for the MCS applications we study hence, they can not be applied by practitioners. Finally, we show how the attacks and defenses of this chapter can be combined in a tool, that we have open-sourced, in order to enable new research on LPPMs for the MCS scenario and the design of privacy friendly MCS applications by developers.
2. Considering the emergence of smart devices in modern households, we study smart meters as it is the new paradigm for measuring electrical consumption. More specifically, as smart metering is a well studied area, we focus on the privacy quantification of one of the most prominent privacy protection mecha-

nisms, that of aggregation. The underlying idea is that if various energy consumptions profiles, from different users, are aggregated and then reported to the energy provider, individuals are protected as the distinguishing patterns in their own profiles will disappear. In this thesis we examine whether aggregated user consumption data offers such privacy guarantees and how the size of the aggregate affect the privacy protection. In order to be able to quantify users' privacy, we first model privacy as a cryptographic game, using an indistinguishability notion. In this game, two single energy load profiles and an aggregated one are given to an adversary, with the goal of selecting which of the two profiles is included in the aggregate. The privacy metric in this case, is how much better an adversary performs than random guessing. The analysis of the privacy for individual households for different aggregation sizes, using real energy consumption data with more than 700 households, shows that an average household is insufficiently protected in aggregates of just a few load profiles. On average, an adversary can distinguish two typical load profiles forming an aggregate with very high ($> 80\%$) probability, when reporting energy consumption information every 15 minutes. Even for larger aggregation sizes, the adversarial advantage is rather high. Furthermore, we examine the influence of various parameters, e.g., temporal resolution, on the detectability of a household within an aggregate. Finally, we show that single energy-hungry appliances can be detected in the aggregates of a handful of households with significant advantage.

3. Finally, considering that users tend to spend significant time on the Internet and knowingly or not their personal preferences are collected on a daily basis, we study users' privacy in microdata collections. The dangers in the specific scenario are well studied and it is demonstrated that with minimal information users can be re-identified in pseudonymized microdata collections [73, 113, 115]. We develop a tool that enables users to approximate their privacy level, given a set of preferences they want to share with a service provider, before actually sharing their data. Hence, users can consciously decide (wrt privacy) whether to share or not their data. The tool works with published statistics about a user's items and hence, does not require access to provider's database. We describe the algorithm running in the tool and the privacy metric that evaluates users' privacy levels. We evaluate the effectiveness of our tool using actual datasets from service providers. Last, aiming to make the tool user friendly, we describe the benefits of visual results against numerical scores, and describe the tool's visual representation.

OUTLINE

Following our contributions in the aforementioned fields, this thesis is split into three chapters. Chapter 2 is about location privacy in mobile crowdsourcing. Chapter 3 is about privacy assessment of aggregation schemes in smart metering. Then, in Chapter 4, we develop a tool for measuring users privacy levels on microdata collections. We conclude this thesis in Chapter 5.

PUBLICATIONS

This thesis consists of work done in the following publications:

- S. Boukoros, M. Humbert, S. Katzenbeisser, C. Troncoso. “On (The Lack Of) Location Privacy in Crowdsourcing Applications.” In: USENIX Security Symposium, 2019
- N. Büscher, S. Boukoros, S. Bauregger, and S. Katzenbeisser. “Two is not enough: Privacy assessment of aggregation schemes in smart metering” In: Privacy Enhancing Technologies Symposium, 2017, pp. 198–214.¹
- S. Boukoros and S. Katzenbeisser. “Measuring privacy in high dimensional microdata collections.” In: International Conference on Availability, Reliability and Security, 2017, pp. 15:1–15:8.

During the course of the Ph.D., the following papers were also published:

- A. Borgwart, S. Boukoros, H. Shulman, C. van Rooyen, and M. Waidner. “Detection and forensics of domains hijacking” In: IEEE Global Communications Conference, 2015, pp. 1–6
- N. Matyunin, N. A. Anagnostopoulos, S. Boukoros, M. Heinrich, A. Schaller, M. Kolinichenko, and S. Katzenbeisser. “Tracking private browsing sessions using CPU-based covert channels.” In: ACM Conference on Security & Privacy in Wireless and Mobile Networks, 2018, pp. 63–74
- A. Taha, S. Boukoros, J. Luna, S. Katzenbeisser, and N. Suri. “QRES: Quantitative reasoning on encrypted security SLAs.” In: arXiv preprint arXiv:1804.04426, 2018
- S. Boukoros, N. P. Karvelas, and S. Katzenbeisser. “A lightweight protocol for privacy preserving division.” In: IEEE Wireless Communications and Mobile Computing Conference, 2017, pp. 717–722
- S. Boukoros, A. Nugaliyadde, A. Marnerides, C. Vassilakis, P. Koutsakis, and K.W. Wong. “Modeling server workloads for campus email traffic using recurrent neural networks.” In: International Conference on Neural Information Processing, 2017, pp. 57–66
- S. Boukoros, A. Kalampogia, and P. Koutsakis. “A new highly accurate workload model for campus email traffic.” In: IEEE International Conference on Computing, Networking and Communications, 2016, pp. 1–7

¹ Parts of the experiments were done by Stefan Bauregger as part of his master thesis.

Published Content

S. Boukoros, M. Humbert, S. Katzenbeisser, C. Troncoso.
 “On (The Lack Of) Location Privacy in Crowdsourcing Applications.” In: USENIX Security Symposium, 2019

Ethical considerations

This chapter includes experiments with actual user data published online. For the work on this chapter, we did not collect any personal data other than that made publicly available by the crowdsourcing projects we study. During our experiments we have limited our inferences to the minimum to prove the privacy threats posed by public data sharing. All data reported is aggregated or anonymized such that no individual’s data is exposed. We have notified the service providers about our findings, and we have provided them with code to implement the defenses so that they can take appropriate measures. At end of this project, all collected and inferred data were deleted. The code from this section will be made available so that it can be used by crowdsourcing applications and improved by the research community. This procedure has been approved by EPFL’s ¹ ethics committee.

2.1 MOBILE CROWDSOURCING

Crowdsourcing is a participative online activity in which the undertaking of a task is outsourced to a group of individuals [34]. This new paradigm of distributing a fragmented task, is an efficient, scalable business model that allows the cheap (or often free) massive collection of data. Indicative of the growth of this data collection methods is the appearance of over 2,000 crowdsourcing platforms[1, 84] in the last years [146]. Furthermore, according to recent industrial reports [58], in the last decade 85% of top global brands have already adopted crowdsourcing, and in 2018 75% of the world’s highest performing enterprises would use crowdsourcing.

As mobile crowdsourcing (MCS), we refer to a special type of crowd sourcing where participants collect geo-located data. Similar to other

¹ Partner institution in this study.

applications where users' data are collected, the goal is to share it with central servers to attain a particular objective. A straightforward use case of crowdsourced geo-located data is mapping, e.g., map generation from volunteered GPS data [2], or map enrichment using geo-tagged photos [3]. Other applications crowdsource the collection of diverse geo-located measurements, such as cellular or WiFi signal quality to localize antennas [4–10], weather data to improve weather predictions [11], or transportation data for accurate real-time traffic information [12, 13]. Other geo-located crowdsourcing applications pursue improving citizens' experience. For instance, in Kamino [14] users publish their favorite routes, and in Stereopublic [15] they report quiet places.

As of 2018, OpenStreetMaps [2], a map generation project from contributed GPS points, reports 4.3 million users,² with 1 million active map editors contributing over 4 billion GPS points. Similarly, OpenSignal [8], a popular network-measuring application, reports over 20 million users.³ Safecast [16], a citizen science project collecting environmental data, currently reports over 75 million measurements from approximately three thousand users. Many other applications are available [3–15], and the ecosystem is growing due to the appearance of platforms aimed at easing the development of crowdsourcing applications [1, 84]. Large IT companies such as Google [17], Microsoft [18] and Mozilla [6] are also beginning to rely on crowdsourcing, e.g., collecting data from billions of users' devices to build WiFi location databases.

Crowdsourcing data collection also enables the provision of scientific or humanitarian services that otherwise would be unattainable due to high cost or lack of sentiment. An example is the Safecast project that provides real-time and historical radiation measurements around the world, valuable for many scientific disciplines and public health. Additionally, many of these projects offer an open-source, free alternative to commercial services. For instance, OpenStreetMaps maps can be used and edited by anyone as an alternative to Google or Apple maps.

MCS can bring great benefits for organizations and society. However, the collection and sharing of geo-located data raises serious privacy concerns. It is known that location data can be used to identify points of interests (POIs) [74, 77, 97], infer users whereabouts and preferences, or de-anonymize anonymous traces [157]. This risk increases when considering auxiliary publicly available information [36, 93, 119], and persists even when protections are put in place [129, 130]. Recent scandals related to the publication of data by fitness applications [19, 20] or irresponsible data analysis by transportation companies [21] exemplify the potential impact of these threats.

² <https://wiki.openstreetmap.org/wiki/Stats>

³ <https://opensignal.com/methodology#over-20-million-users-of-our-app>

Over the last decade the research community has proposed a vast number of location privacy-preserving mechanisms (LPPMs) to address these issues [128], and can even provide strong differentially private guarantees [33, 45, 70] and even offer optimal utility [44, 122]. Even though it seems like the location privacy question is technically solved, the reality is that these LPPMs *solely focus on one use case*. They are generally geared towards location-based services (LBSs) in which whenever users need a service, they reveal their location to a provider (e.g., to find nearby restaurants). In this context utility is user-centric and hinges on the precision of the reported locations. In MCS applications, on the contrary, geo-located data is often shared continuously and over long periods and, while the data utility is still correlated with the location precision, it is foremost tied to the values of the measurements reported at these locations (e.g., WiFi signal strength, or radiation level). Moreover, MCS utility cannot be captured with a user-centric approach as, by definition, MCS benefits from aggregating data collected by a large amount of users.

Chapter outline

The rest of this chapter is organized as follows. We first discuss related work in the field of location privacy in Section 2.2. We subsequently introduce the three datasets used throughout this chapter in Section 2.3. Then, in Section 2.4 we validate the privacy threat of publishing raw continuous location data. Section 2.5 describes the LPPMs tested and the privacy and utility metrics. In Section 2.6 we present the experimental results while in Section 2.7 we discuss how the this chapter’s results and code can be used to enable new research on privacy preserving MCS, and best practices for developers.

2.2 RELATED WORK

We now present related work regarding the uniqueness of users’ trajectories, ways to extract points of interest from trajectories, location privacy attacks, defenses and privacy quantification. We refer the reader to the following surveys for further information about the security and privacy landscape of location data [79, 98, 128, 148].

2.2.1 POIs and the uniqueness of human mobility

Given a collection of a user’s location reports, one is able to infer where this person frequents. We present some of the most widely used strategies for POIs extraction in the next paragraph. Furthermore, we discuss various works that illustrate why human mobility can be used as identifier.

2.2.1.1 *Extracting POIs*

Extracting users' POIs, from a collection of location data, is a first step for enabling various other attacks. Traditionally, POIs are extracted using machine learning techniques, a paradigm we also follow in this chapter. Cho [50] used machine learning to infer users' POIs and predict their next destination. For the POIs extraction relied on clustering techniques and decision trees while for the prediction part on Markov models. Hoh et al. [87] aimed to enhance security and privacy in traffic-monitoring systems. As part of their work they conducted a case study to infer users' home addresses from collected geo-located data. They relied on the K-means algorithm to create clusters and then using heuristics they were able to detect individuals' home locations. Urner et al. [151] studied the effects of various input features to machine learning algorithms for next place prediction. In order to extract the places a user stayed for a least a certain duration, i.e., their POIs, they also used K-means clustering.

2.2.1.2 *Uniqueness of human mobility*

Several studies have examined individuals' trajectories and their uniqueness. Gonzalez et al. [80] studied anonymized mobile phone data. They explored the statistical properties of the population's mobility patterns such as the distance between consecutive calls or returning to a specific place. Their results indicate that human trajectories have a high degree of temporal and spatial regularity. Similarly, studying anonymized phone data, Song et al. [142] explored to which degree human mobility is predictable. Their results highlight that individual's trajectories are predictable despite of the distance users cover. In other words, an individual's location data history is a unique identifier. In the same direction, de Montjoye et al. [57] investigated how the uniqueness of mobility traces decays depending on their resolution. They show that uniqueness cannot be avoided by lowering the resolution of a dataset. Zang et al. [157] analyzed call records from millions of users in the US where they extracted users' top POIs. Their results illustrate that anonymization of location data does not work as the top-N POIs per user have very small anonymity sets.

2.2.2 *Privacy mechanisms*

A commonly used privacy-preserving mechanism is the perturbation of the actual location with noise. The state of the art in spatial obfuscation is geo-indistinguishability [33]. This mechanism adapts differential privacy notions to location data, offering privacy guarantees independent from the adversary's prior information. It is known however that the application of independent noise to continuous releases of correlated locations leads to privacy loss. Chatzikokolakis

et al. [45] tackled this problem by proposing a more efficient mechanism. Their proposed mechanism examines if continuous releases of locations are correlated and decides whether new noise should be applied to the releases. In order to enhance the utility of the perturbed locations, Chatzikokolakis et al. [44] proposed a mechanism called optimal remapping. Based on prior information about users' locations, the newly reported locations are being re-mapped to places that offer the same privacy guarantees but higher utility. This remapping takes into account both the prior probability of all nearby locations as well as, the distance from the perturbed location.

Hoh et al. [88] proposed a LPPM based on hiding locations. The algorithm uses information from all users in the system to decide if individual locations should be revealed. Furthermore, the system uses a novel time-to-confusion privacy metric, specifying how long an individual can be tracked. Huang et al. [89] developed a scheme to both maintain quality of service, defined by the provider, and privacy. In their system, users' mobile devices have two states, one in which they maintain a communication identifier when transmitting data and one that is silent. They exploit the time in between switching rounds, and hence communication identifiers, to provide privacy by breaking the link between previous and current identifier.

Generalization techniques have been proposed in the context of location privacy. Bamba et al. [37] developed a framework for anonymous location-based queries to LBS providers. Their system operates on a discretized map and users can specify their desired privacy level (defined by k -anonymity or l -diversity) and the temporal and spatial resolution. The framework then receives users' requests, further anonymizes them and then forwards them to the cooperating LBS providers, which need to have a part of the system installed on their side. Gruteser and Grunwald [83] also developed a centralized framework for location privacy. The underlying idea is that a desired degree of anonymity, measured by k -anonymity, can be achieved by generalizing reported location, as the reported wider area will include further $k-1$ users.

Generative models, from actual user data, have also been proposed as LPPMs. Chen et al. [47] explored the use of differentially private N -grams, i.e. a probabilistic model from sequential items for next item prediction. Their generative model takes as input actual sequential user data. Then, creates a differentially private variable N -gram model by adaptively adding noise to the model's levels. Acs et al. [29] explored privacy preserving spatio-temporal density releases. Using actual user trajectories they keep a sample of locations per user and then they perturb the time series to preserve privacy. Finally, they released privacy-preserving statistics for the city of Paris. Bindschaedler and Shokri [41] proposed a generative model for creating synthetic locations. Their model extracts geographic and semantic information

from actual traces in order to achieve low utility loss. Furthermore, the privacy of trajectories is maintained as the original seeds that generate the traces have been crafted in such a way that maintain users' indistinguishability.

2.2.3 *Attacks on geo-located data*

Various attacks in the literature have highlighted the danger of sharing raw geo-located data. Krumm [98] studied collected location data from volunteers. Using heuristics and clustering techniques he was able to retrieve a large amount of subjects' home addresses. Furthermore, by reverse geo-coding those addresses, he was able to retrieve some of the subjects names. Using three defenses, spatial cloaking, simple noise addition (not geo-indistinguishability) and generalization he further explored how much obscuration is required to maintain users' privacy. Freudiger et al. [74] analyzed the privacy threat by LBSs. They identified various ways that individuals share location data with providers and several attacks (e.g., identification of work/home locations, extract POIS etc.) that the provider can carry out on the data. Their experimental results highlight that even when users share a small amount of location data, they can be uniquely identified. Gambs et al. [77] explored the re-identification risks in location data. Using clustering techniques they extracted each users' POIs, from a set of raw location data, and used them as states for Markov models. Furthermore, they also calculated the probability of transiting from one state to another. Then, they were able to build de-anonymizers for anonymous datasets. Similarly, using hidden Markov models, Mathew et al. [112] show how an adversary, by having access to geo-located data, can predict users' future movements.

Besides the geo-located data someone shares, auxiliary sources of data can de-anonymize individuals or enable inference attacks. Khazbak and Cao [93] explored if co-location information can de-anonymize mobility traces. Using two location datasets, they developed models based on maximum likelihood estimation (MLE) and hidden Markov models to infer users' past locations in anonymized datasets, based on available information from other co-located people. Backes et al. [77] studied whether location data can reveal social relationships. Using machine learning they show that individuals' traces are more similar to those within their social circles than strangers. Furthermore, they explored if and to what extend various countermeasures can be applied to mitigate such inference attacks.

In the front of countering proposed location privacy defenses, Pyrgeles et al. [130] evaluated the effectiveness of aggregation in geo-located data. Using a game-based privacy metric (similar to the one we propose in Chapter 3) over real user data, they investigated both the effectiveness on raw and differentially private datasets. Their work

highlights that aggregation does not offer strong privacy guarantees. Moreover they studied various parameters affecting the privacy of the aggregate such as density of observations or timing. Last, they show that large amount of noise could be introduced to protect users' privacy in the aggregate however this has detrimental results on the utility of the data. Xu et al. [156], also focusing on aggregation, show that an adversary can reconstruct individuals' trajectories using only the aggregated data. They developed an unsupervised framework that exploits the uniqueness of human trajectories for different parts of the day and achieved up to 91% recovery accuracy in their experimental results over actual user data.

2.2.4 Privacy quantification

Shokri et al. [139] proposed a framework for evaluating LPPMs and a novel privacy metric based on distortion. This measure captures the adversary's correctness regarding users' original traces when LPPMs are in place. Shokri et al. [140] also developed a framework to quantify location privacy. The platform supports well known attacks and defenses and privacy is defined as the adversarial correctness. Oya et al. [122] studied the design of optimal LPPMs. Their results illustrate that even if mechanisms are optimized over the expected adversarial error, there are aspects of privacy that are not taken into account. Hence, they proposed the use of conditional entropy and worst-case loss as auxiliary privacy measures.

Our contribution

Contrary to previous works, in this chapter we propose a framework to quantify privacy in the context of mobile crowdsourcing. In our analysis, we test the effectiveness of the geo-indistinguishability defense, and two of its variants. Moreover, we test location hiding techniques and a generalization method. Similar to previous work, we use an automated way to extract users' POIs and hence, infer sensitive information about them. However, the proposed privacy metrics are novel, easily understood by developers and users, and are better suited to the crowdsourcing scenario where users release streams of their location data. Finally, we rely on realistic utility functions to evaluate the privacy-utility trade off in the datasets under study.

2.3 MCS APPLICATIONS UNDER STUDY

In this section, we introduce the two crowdsourcing applications studied in detail in this chapter. In Section 2.4.2 we also briefly investigate OpenStreetMaps [2] to assess the generalizability of our results.

2.3.1 *Safecast*

Safecast [16] is an international, volunteer-centered, organization whose goal is to monitor the global radiation levels and detect abnormalities in near real time. Safecast crowdsources the collection of radiation data. It provides users with open hardware devices that collect radiation measurements every 5 seconds. The measurements are published in three ways: (i) as raw measurements available through an API, (ii) as curated measurements available as a bulk download and (iii) visualized as an interactive map.

2.3.1.1 *Safecast API dataset*

Safecast provides an API to access raw radiation measurements. Radiation measurements contain the user’s name, a unique user ID, the device’s ID, latitude and longitude, a UTC timestamp, and the radiation value and units. No registration is required to access this data and Safecast’s privacy policy reads⁴: “All data collected by Safecast is released under a CCo public domain designation. Anyone is free to use with no licensing restrictions. We have done this to enable the most flexibility in it’s use by others”.

Using this API we retrieved data for all users who had at least 30 measurements. We obtained almost 64 million measurements collected between 2011 and 2017, from 551 users. Many users provide their actual name (often including their surname) or use identifying nicknames, and some add metadata to the measurements explaining the purpose of the trip in which the data were collected. Before using this dataset in our study, we dropped all entries in which the latitude, longitude, timestamp or the username were empty. We also removed data uploaded by several users belonging to one organization under a unique ID (e.g., corporations or universities), and we converted all UTC times to the local time determined by the measurement’s coordinates. After this process, the dataset has 52.8 million measurements from 539 users.

2.3.1.2 *Safecast curated dataset*

This dataset, available as a bulk download, is the source of Safecast’s interactive map. The curation steps, involving sanity filters and manually-maintained exceptions, are explained on Safecast’s website.⁵ It contains 64.2 million measurements from 608 users, collected from 2011 to 2017. We removed IDs corresponding to organizations, malformed entries, and converted all UTC times to local. After this process, the dataset has almost 56.7 million measurements from 537 users (534 of which also appear in the API dataset).

⁴ <https://blog.safecast.org/faq/licenses/>

⁵ <https://safecast.org/tilemap/methodology.html>

2.3.1.3 *Safecast utility*

The Safecast project uses the collected data to study different phenomena related to radiation. We consider two of the possible uses of the data.

First, we consider the visualization of radiation on the interactive map. Safecast computes the visualized radiation levels from the crowd-sourced measurements. For a given region of interest, the map values are obtained as follows. First, Safecast filters the measurements within the region and computes the average radiation at each location over the last 270 days. Second, they discretize the area to 2.25 million grid points (1500 discrete locations per axis). The displayed map is created using nearest-neighbor interpolation on the points of the grid using the averaged radiation measurements. The reported radiation is measured in counts per minute (cpm), expressing how many ionized particles are detected per minute by a monitoring instrument. This use case represents a scenario in which the measurements processing required to obtain the goal (the interactive map) is in principle amenable to noise.

A second case of interest, related to Safecast's concern about public safety, is the detection of 'hotspots' – specific areas where radiation is above a pre-defined threshold. These 'hotspots' indicate locations where radiation could be harmful. Once identified, Safecast might send experts to perform on-site examination to better understand the causes and consequences of such dangerous measurements. Therefore, it is important that the localization of hotspots is rather exact in spatial terms.

2.3.2 *Radiocells*

Radiocells [5] is an open-source community project whose goal is to provide an open-source alternative to commercial, closed source, network geo-location databases. Additionally, they aim to provide raw data on telecommunication infrastructure that can be use for diverse experiments. Radiocells crowdsources the collection of measurements via a mobile application called 'Radiobeacon'.⁶ This application allows users to take continuous measurements as they perform daily activities. Users can choose when to start measuring and when to stop, and when to upload the measurements to the Radiocells server. Furthermore, they can select a specific area where measurements will not be recorded, e.g., to protect their home locations from appearing in the dataset. However, previous work shows that this kind of defense is rather fragile [86].

⁶ <https://f-droid.org/packages/org.openbmap/>

2.3.2.1 *Radiocells dataset*

The raw data uploaded to the server is publicly available for download. The data is licensed under Creative Commons Attribution-ShareAlike 3.0 Unported and ODbL licences aimed at not restricting the use of the data.⁷ Amongst other information, measurements include: signal strength, cell (antenna) ID, location, timestamp, and smartphone’s model, software, OS version, and manufacturer. Contrary to Safecast, this dataset does not contain usernames in an effort to preserve users’ privacy. However, the combination of the smartphone characteristics, the location, and the network provider is likely to represent a quasi-identifier. We downloaded data for 2013 to 2017, obtaining 25 million measurements. To separate users, we grouped the measurements according to phone manufacturer, phone model, country and network operator. We dropped cases where we observed a tuple (country, manufacturer, model, operator) with multiple revisions of the operating system, as we cannot distinguish OS up/down-grading by a user from multiple users with a different configuration. We obtained 998 potential unique users, of which we only kept those that had more than 100 measurements. We inspected the remaining measurements for spatial inconsistencies. To this end, we calculated the required speed to travel between contiguous measurements. We removed users whose maximum speed was more than 200 km/h (i.e., faster than a train). The final dataset contains 568 users and about 4 million measurements.

2.3.2.2 *Radiocells utility*

The Radiocells database, amongst other purposes, can be used for antenna or device geolocation. Contrary to Safecast, Radiocells does not provide explanations or code as to how they produce their map of antennas. Thus, we use the location function described by OpenCellID [4], another crowdsourcing project aimed at geo-locating antennas. This function defines the location of an antenna as the average of the latitudes and longitudes of the measurements referring to this antenna. Such results are useful to enable scientific studies about the signal quality in specific places and the distribution of antennas. Furthermore, the Radiocells database can be used (online or offline) for device geolocation [22]. Users can query either a local instance of the database or Radiocell’s online service to find their coordinates based on the nearest antenna and their device’s signal strength.

2.4 PRIVACY RISKS IN CURRENT MCS APPLICATIONS

In this section, we assess the privacy risks associated to existing crowdsourcing applications data publication approaches using well-known inference techniques. We study both Safecast and OpenStreetMaps,

⁷ <https://radiocells.org/license>

Table 2.1: Safecast API dataset statistics.

Measurements	Users	Avg measurements	Distinct days
<10k	213	3331	5
10k-100k	230	38341	20
100k-1M	87	270,387	105
>1M	10	1,958,760	632

since they provide identifiable information about their users, enabling us to validate our inferences against information available in other online platforms. We do not perform any attacks on the Radiocells dataset, as it did not include users’ actual names and hence, we were not able to collect ground truth. However, this by no means indicates that users are sufficiently protected. It is known that if anonymized or pseudonimized datasets are correlated with auxiliary sources of information, then they can be de-anonymized [113, 115] (we elaborate on this issue in Chapter 4). We stress that our exploration barely scratches the surface of what can be learned from the published data. A determined adversary with enough resources could exploit the published data sources to infer much more information about users.

2.4.1 Safecast

We evaluate privacy risks for Safecast users on the *Safecast API dataset* (see Section 2.3.1). This dataset contains usernames and raw geo-located measurements. We split the users in the dataset into four groups according to the amount of reported measurements they have. For each group, Table 2.1 shows the number of users, their average amount of measurements, and the average number of days in which they took at least one measurement. From each group we select the 10 users with the most measurements that provide their real names as targets for inference. Since in the group with the most measurements there are only 4 users with real names, we end up with 34 target users. This allows us to manually validate our inferences in reasonable time.

2.4.1.1 Identifying POIs

Inspired by previous works which show that clustering is a necessary step to extract POIs [50, 69, 74, 87, 97, 112, 151] we rely on the density-based clustering algorithm (DBSCAN) [69] to find work places. Contrary to other clustering algorithms (such as K-Means), DBSCAN is robust to noise and outliers and does not require to specify the number of clusters a priori. The algorithm receives as input all locations (also referred to as points) reported by a user, the minimum required amount of points per cluster, and the maximum allowed distance

between the cluster's points. DBSCAN starts by randomly selecting a point c . Then, it finds all points p that are in distance ϵ from this point. Then, from the points p reachable from the first point, it tries to find more points q where q are reachable directly from p but not from c . If at the end of this procedure the minimum points have not been reached, it moves to another random point and starts all over again. In order to use the reported locations which are in tuples of latitudes and longitudes, we converted all distances to radians first. Moreover, we used a ball tree data structure to speed up the neighbors queries. The algorithm outputs a label for every location, indicating to which cluster it belongs, or if it has been labeled as noise.

2.4.1.2 Identifying workplaces

We run DBSCAN on every users' measurements during working hours (Monday to Friday from 9AM to 5PM), and we keep the five clusters with the highest number of points.⁸ In many cases these are rather large clusters with too many POIs. To ease manual validation we use X-means clustering [125] to split the large clusters, and consider as POI the centroid of the two largest subclusters. We end up with at most 10 POIs per user. We use the MapQuest API [23] to obtain these locations' addresses and, if existing, the names of businesses at those coordinates. Note that we could have collected all POIs in the clusters (such as corporations, parks, restaurants etc.) and filtered out those not related to workplaces (manually or automatically).

The size of the clusters depends on how DBSCAN is configured. We consider two parameter settings: a *tight* configuration that attempts to collect strong evidence of a POI before proceeding to manual check, and a *loose* configuration which provides more candidates but many false positives to be filtered manually. The *tight* configuration requires that there are many points in a cluster (at least 75), that these points are very close (30 meters, roughly the size of a small building), and that the user spends significant time there (at least 30 minutes per day with measurements) to avoid false positives related to commuting patterns. The *loose* configuration does not consider the 30-minutes constraint and relaxes the number and distance between points in an adaptive manner. Starting from 80 points separated by 60 meters, if no clusters are found we increase the distance by 30 meters (up to 120 meters maximum) and decrease the number of points by 15 (down to 35 points).

Using the tight configuration, we recover and validate the workplace in 7 cases out of the 34 (21% recovery rate). The loose configuration can reach a recovery rate of 35%. For validation we relied on professional social networks such as LinkedIn or the users' personal webpages. For 9 non-validated cases the users did not have a page or had too

⁸ Throughout this chapter we use the terms points and locations interchangeably.

common names to find their correct information. We consider the results for these users not to be necessarily wrong but inconclusive.

If we look at the different users' groups, in the loose configuration we find the workplaces for 40% of the users with less than 10k measurements, 20% of the users with 10k-100k measurements, 50% of the users with 100k-1M measurements, and 25% of the users with more than 1M measurements. This shows that, surprisingly, the amount of shared data does not seem to be correlated to the privacy risk. On the contrary, it seems to be highly dependent on the collection patterns of the users. Generally, we observe that people fall in two categories. Those who travel to specific places in order to take measurements, and thus their work/home addresses cannot be inferred, and those who measure radiation in the area they live in. The Safecast co-founders, who are the top contributors in terms of data points, fall in the first category, explaining the low inference power for users with more than 1M measurements.

Our results confirm recent findings in the literature regarding personal information inferences from location data[61, 66]. Yet, we want to stress that the threat may be worse for MCS, due to the volume of data exposed by participants. For reference, Safecast's lowest contributing group has on average 3k measurements per user (see Table 2.1) while in the Twitter analysis performed by Drakonakis et al. [61] only the top contributing users (less than 0.06%) have more than 3k geolocated tweets. Thus, even if the number of MCS users is not as large as social networks' users, we expect a significant fraction of them to be vulnerable to attacks.

Finally, we note that the results are also limited by our manual validation effort. If we take all POIs in a cluster as candidates, instead of taking the centroids, the chances of identifying sensitive POIs increase, but so does the cost the attack. For instance, using the API of OpenStreetMaps⁹ we can see in Figure 2.1 that using the tighter configuration, we find clusters covering less than 0.5 km² containing less than 10 POIs in total, whereas in the looser configuration we find larger regions containing between 10 and 200 POIs. Larger clusters contain more POIs to be examined, which increases the possibility of false positives. Nevertheless, we note that the semantics of locations often make it easy to filter out false positives, e.g., lakes or parks can be usually discarded as workplaces.

2.4.1.3 Other POIs

A deep analysis of the times and semantics of the POIs identified by DBSCAN can reveal other sensitive information about users. Among others, we could infer two users' membership to specific organizations: one member of the Scientology church who reported many points from the Church of Scientology Celebrity Centre in a major city; and a

⁹ https://wiki.openstreetmap.org/wiki/Overpass_API

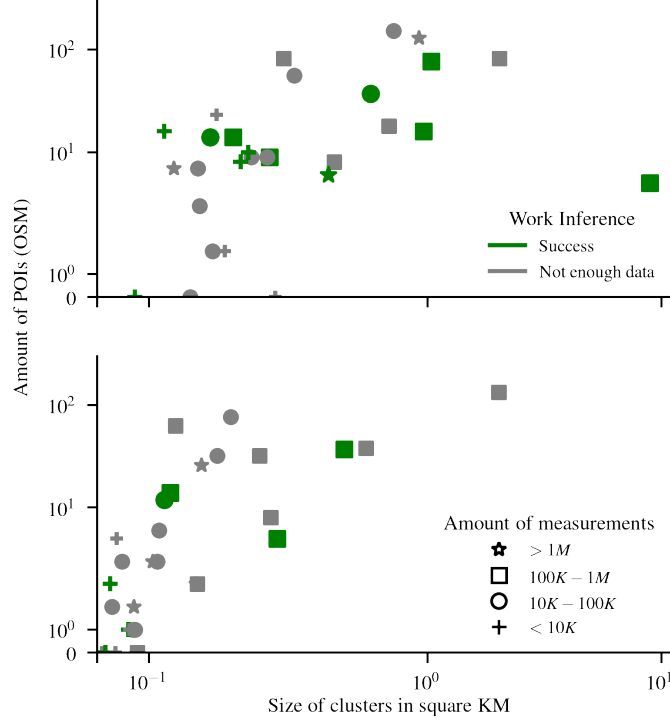


Figure 2.1: Cluster size and amount of POIs per user. Looser setting – 35% success (top) and tight setting – 21% success (bottom).

Masonic lodge member who regularly visited the lodge headquarters. We could verify this information online for both users. Second, we identified two work-related activities: a US-based scientist working on a paper regarding radiation around a lake in the Southern part of the US, and a photographer working in a Japanese city. We validated the former inference using Research Gate, and the latter on the webpage of the artist. Finally, we could also follow the education steps of a European PhD student. Her points of interest over time reveal the university where she obtained her master’s degree, an exchange with another European university in the framework of the Erasmus project, and the university where she is completing her doctoral studies. We verified these facts on her CV available online.

2.4.1.4 Inferring social links

Knowing from previous work [54, 64] that co-locations can unveil social links, we examine whether crowdsourced geo-located measurements can also be used for this purpose. We identify pairs of users with similar latitude, longitude, and time of measurement in the whole Safecast API dataset. More advanced methods, such as measuring the amount of time two users are co-located or the number of different locations where two users jointly report their locations [36, 54, 153], could further improve our results.

In total, we identified 50 unique pairs of users with real names (last and first names) *and* with at least one co-located instance. Among them, 16 are real friendship relations according to public information available on online social networks (Facebook and Twitter), i.e., 32% correct inferences. Note that, in order to validate a social link, we need both users in a pair to be part of the same social network, and that at least one of them publicly reveals her social links. The pairs we could not validate did not fulfill one of these conditions. We cannot conclude that these pairs of users are not socially related. In fact, given the sparsity of the dataset, there is a high likelihood that they know each other.

2.4.2 *OpenStreetMaps*

As cross validation of the results in Safecast, we attempt to repeat the inferences on data collected from OSM. Contrary to Safecast, OSM does not have an open API for accessing users' data. Yet, traces from those users who have chosen to make their data available can be easily obtained from OSM's website. To minimize the impact on OSM servers, and comply with their non-crawling policy, we manually downloaded data for 30 users with a large amount of contributions¹⁰ of which 17 used their actual names (or indicative nicknames). Although the majority of the points in the dataset were rather old (most of them at least 7-8 years old), we were able to verify previous workplaces for 3 of the 17 users (17%). For other users, we found out that they did not have a standard place of employment during data collection period (e.g., students). Additionally, for *all* users, their POIs were within the area where they worked or lived. We used this fact to infer two of the users' short vacation trips which we manually verified with information publicly accessible from their social media accounts.

2.5 PROTECTING LOCATION PRIVACY IN MCS

In the previous section we showed that current data sharing practices of MCS applications put users' privacy in danger. In this section, we describe the existing LPPMs that could be put in place to reduce the privacy risks. These defense mechanisms have been proposed for the case of LBSs which is fundamentally different from MCS. First, LBSs aim at fulfilling an individual, user-centric, need related to a user location (e.g., find nearby restaurants). MCS, aim at fulfilling a common objective through collaborative measurements. Second, LBSs often require sparse geo-located data (just few points per geo-located query) where MCS usually requires continuous collection. Therefore, it is not straightforward that these LPPMs will work as expected in the MCS setting.

¹⁰ <http://resultmaps.neis-one.org/oooc>

2.5.1 Defenses

We consider the following three main LPPMs categories [98, 140]: (i) spatial obfuscation, (ii) hiding, and (iii) generalization. We do not consider the use of dummy locations or synthetic data [41, 47], as we consider these solutions to be inadequate for the MCS setting. These approaches focus on producing plausible locations, but to the best of our knowledge there is no proposal that provides the means to generate measurements (or other values) to be associated to the reported locations while preserving utility in applications such as those in our study. In fact, we argue that generating fake measurements, possibly using prior information, is bound to pollute the real-time measurements that these applications aim at collecting.

2.5.1.1 Spatial obfuscation

As already discussed, the state-of-the-art spatial obfuscation technique is *geo-indistinguishability* (*GeoInd*), and is widely used in the literature [24, 70, 92, 108, 127, 155]. Following the original definition in [33] we obfuscate locations by adding planar Laplacian noise. The magnitude of this noise is controlled by the parameter $\epsilon = l/r$ which guarantees that the ratio between the probabilities of two different locations being the real location in an area of radius r is at most l .

Release-GeoInd. A known problem when using *GeoInd* for continuous reporting is that the level of privacy decreases linearly with the number of reported locations. To address this limitation we consider a mechanism that, inspired by the predictive approach proposed in [45], reports a new noisy location if and only if the user has moved at least z meters away from his previous location. Otherwise, it repeats the last reported location. We call this approach “Release-GeoInd”.

GeoInd-OR. It has recently been shown that remapping the obfuscated locations to the popular ones, according to prior knowledge on users’ movements patterns, can offer optimal utility without reduction in privacy [44, 122]. We implement Chatzikokolakis et al. [44] remapping approach to remap *GeoInd* samples. We refer to this approach as “GeoInd-OR”.

2.5.1.2 Hiding

This defense achieves privacy by not reporting some of the users’ locations [88, 89]. The released locations are *not* perturbed. We consider two hiding strategies: (i) a “Random” strategy in which users release a random subset of their points, and (ii) a “Release” strategy in which users only reveal a new point when they have traveled at least x meters away from the previously reported location.

2.5.1.3 Generalization

This defense reduces the precision with which locations are reported [37, 83]. Similarly to Krumm’s discretization mechanism [98], we implement this approach by reducing the precision of the reported GPS coordinates. We denote this defense as “Rounding”.

2.5.2 Measuring privacy

Location privacy metrics in the literature are mostly based on a function of the distance between the real location of the user and the one inferred by the adversary [139, 140]. This function could be the Hamming distance when measuring semantic similarity, the Euclidean distance when quantifying the *correctness* of the adversary’s inference [140], or entropy when measuring the uncertainty of the adversary regarding the user’s location [122]. These privacy metrics are very well suited for the case of LBSs, where users release one location per query, and the adversary tries to infer users’ location. However, they are hard to use in the MCS setting, where the adversary has access to continuous releases by the user over several days. Thus, it is not possible to establish between which points should distance be computed, or across which points compute a probability distribution for entropy-based metrics.

Furthermore, we argue that none of previously proposed metrics capture privacy in a tangible manner understandable by users and developers of crowdsourcing applications. How much privacy is an error of 10 meters or 500 meters? It is clear that one is larger than the other, but not how much privacy they provide regarding the potential inference of sensitive information. Even more complicated is the case of entropy, whose units of measurement – bits, nats, or hartleys – are rarely known, let alone interpretable, by layman people.

2.5.2.1 Privacy gain

Aiming towards user empowerment (both by users but also application developers), we propose to quantify privacy as the loss of adversarial inference power regarding two privacy dimensions understandable by users: geographical area and POIs. To quantify this loss we use two well-established statistical measures: precision and recall. The former captures the case where privacy is increased when, after the defense, the adversary identifies many false candidate locations along with the user’s real whereabouts, i.e., the adversary has low *precision* ($\frac{TP}{TP+FP}$, where TP and FP refer to *true positives* and *false positives*, respectively). The latter captures the case when, after the defense, the adversary cannot correctly identify the original locations visited by the user, i.e., the adversary has low *recall* ($\frac{TP}{TP+FN}$, where FN refers to *false negatives*).

Spatial privacy gain. Spatial privacy refers to the geographical area in which the adversary infers the user can be (defined as the union of the clusters generated via a POIs extraction attack, as in Section 2.4). We define the true positives (TP) as the area in the intersection of the clusters' union before and after the defense is deployed (i.e., the proportion of the area inferred by the adversary that corresponds to the user's real location). Similarly, we define the false positives (FP) to be the area in the clusters after the defense that does not overlap with the area before the defense (i.e., the area inferred by the adversary where the user was not present), and false negatives (FN) as the area before the defense that does not overlap with the region inferred after the defense (i.e., the area where the user has been but that is missed by the adversary).

POI privacy gain. In reality though, the geographic area itself may not fully capture users' privacy [139]. For instance, if there is only one point of interest in a large area, privacy should be low. On the contrary, in small areas with many POIs (e.g., a block in a city), privacy might be large. To account for this fact we propose a complementary metric based on POIs. In this case true positives (TP) are the POIs in the intersection between the clusters before and after the defense is applied. Similarly, false positives (FP) are POIs identified after the defense that were not present before, and false negatives (FN) are the POIs in the original clusters that do not appear after deploying the defense.

2.5.3 Measuring utility

Besides privacy, one key aspect to decide which LPPM is best for a use case is the LPPM's utility loss. We now introduce the utility metrics used in our evaluation.

2.5.3.1 Distance-based metric

In this chapter, we refer as distance-based metrics those proposed in the literature for LPPMs in the context of LBSs. In our experiments we use the per-location distance between original and obfuscated locations using the haversine distance, necessary for correctly calculating the distance between two points on a sphere given their longitudes and latitudes. The haversine distance is expressed as:

$$d = 2R \arcsin \left(\sqrt{\sin^2 \left(\frac{\phi_2 - \phi_1}{2} \right) + \cos(\phi_1) \cos(\phi_2) \sin^2 \left(\frac{\lambda_2 - \lambda_1}{2} \right)} \right)$$

where R denotes the radius of the earth (approximately 6371 km), ϕ_1 and ϕ_2 the latitudes of the first and second point respectively while λ_1 and λ_2 the longitudes of the two points. We note that, when utility

Table 2.2: Safecast (top) and Radiocells (bottom) measurements per region. Vulnerable users are those with at least one cluster.

Region	Users	Meas/ments	Average per user	Standard deviation	Vulnerable users
Tokyo	30	2,701,367	90,046	203,576	24 (80%)
Fukushima	104	7,765,773	74,671	260,671	65 (62%)
World	537	56,655,768	105,504	70,954	349 (65%)
World	568	3,710,547	6,532	17,312	91 (16%)

is measured as the expected distance, then it can be seen as the dual of privacy, i.e., the further from the real location the obfuscated location is, the more difficult is the inference but also the more utility is lost.

2.5.3.2 Aggregate statistics

In the case of MCS, the utility is inherently collective as it depends on multiple users' measurements. Indeed, most MCS providers are interested in aggregate statistics computed over individuals' contributions. This is the case for both Safecast, to obtain the radiation map, and Radiocells, to obtain the coordinates of the antennas. Both are derived from average measurements of MCS users. Hence, we consider a second utility metric that is not a distance, but a function of the values of interest for the MCS provider using the actual utility functions of the projects (as described in Section 2.3).

2.6 EXISTING LPPMS PERFORMANCE IN MCS SCENARIOS

In this section, we evaluate the performance of the existing LPPMs in terms of privacy gain and their impact on the utility of the MCS applications we study.

2.6.1 Experimental setup

We run the experiments on all the data available from Safecast and Radiocells. To understand the impact on utility, for Safecast we also consider two regions in Japan with very different radiation profiles: Tokyo, where the radiation profile is quite uniform, and Fukushima, where the nuclear incident at the Daiichi power plant [25] in 2011 created areas with elevated radiation.

Table 2.2 summarizes the statistics (number of users, total amount of measurements, and measurements per user) of the regions under study. We report the percentage of users vulnerable to our attacks *before* the defenses are applied, i.e., the percentage of users for which we can find at least one cluster. For Safecast-Tokyo we only report the

statistics of the 30 users considered for protection by the GeoInd-OR (see below).

For a given LPPM we evaluate privacy gain and utility loss as follows:

Step 1. We first obtain clusters for all users in the region of interest from the raw data. We use the tight configuration described in Section 2.4. To identify the POIs, instead of considering just the centroids we use the OSM API to *collect all POIs within the union of clusters of the targeted user*.

Step 2. We apply the LPPM to all users' data and obtain the clusters and POIs from the obfuscated data as in Step 1. Note that when Rounding to 2 or 3 digits, obfuscated locations are separated by approximately 1,100 meters and 110 meters, respectively. In this case the tight clustering parameters with a maximum distance between points of 30 meters will not find any clusters, and increasing this distance would result in too large uninformative clusters. However, the operation of Rounding guarantees that given an obfuscated point, actual location of the user is within a square of size 110, resp. 1,100 meters, centered in the reported location. Thus, for this case instead of using DBSCAN clusters, we pick the squares around the five most frequently reported obfuscated locations.

Step 3. We measure the privacy gain as described in Section 2.5.2. For the Spatial privacy gain, we compare the area (in square kilometers) of the clusters before and after the LPPM, while for the POI privacy gain compare the number of POIs within the clusters.

Step 4. We measure the utility loss as described in Section 2.5.3. In the case of aggregate statistics the utility is application dependent. For Safecast, we consider as utility loss the difference between the radiation values on the application's interactive map (see Section 2.3.1) before and after the LPPM. We consider the absolute difference in cpm per grid point. In Radiocells, we consider as utility loss the distance between the location of the antennas before and after the LPPM.

2.6.2 Defenses implementation

For the GeoInd defense, we set the privacy parameter $\epsilon = \ln(1.6)$, and use a radius of $r \in \{50, 150, 300\}$ meters, which yields $\epsilon \in \{0.01, 0.003, 0.001\}$. The noise is drawn by first transforming the location to polar coordinates. Then, the angle is drawn randomly between 0 and 2π while the distance is drawn from

$$C^{-1}(\rho) = -\frac{1}{\epsilon} (W^{-1}(\frac{\rho-1}{e}) + 1)$$

with W^{-1} denoting the -1 branch of the Lambert W function [53]. Finally, the generated distance and angle are added to the original location. Figure 2.2 present the CDF of the noise added on all GeoInd

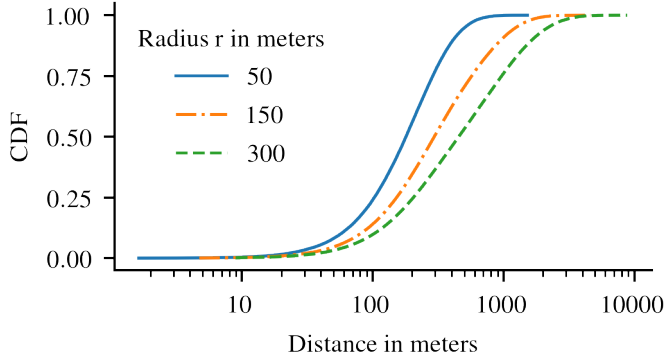


Figure 2.2: GeoInd noise magnitude for different radius ($l = \ln(1.6)$).

variants (Optimal Remapping and Release-GeoInd). The noise is controlled by either the radius (r) or the privacy parameter (l). For the Release-GeoInd mechanism, we use $l = \ln(1.6)$, $r = 50$ meters, and we select the distance between released locations to be $z \in \{30, 60, 90\}$ meters. Remapping the locations for the LPPM GeoInd-OR requires computing the posterior probability for every candidate location. This operation is rather costly when the number of locations being considered grows. To keep a reasonable experimentation time, we only test GeoInd-OR for the Tokyo region in the Safecast dataset. We use 80% of the users to construct the prior probability distribution describing users' movements, and the remaining 20% to evaluate the effectiveness of the approach. We chose this 20% manually to keep a balanced testing set. It is composed of the top 10 users with many (more than 50k), moderate (between 10k and 50k), and few (less than 10k) measurements.

Optimal Remapping. For the optimal remapping technique we follow these steps; For performance reasons, we first round each location to 3 digits, in order to merge nearby locations together. Then, we calculate the probability of each coordinate. Afterwards, we convert all coordinates to a Cartesian system using their distance from the center of the Earth. A useful tutorial on this can be found in [26]. Using the Cartesian coordinates we build a KD-Tree for efficient nearest neighbor calculations. Then, for every location where GeoInd has been applied, we query all nearest neighbors in a region r' . This r' is set to be as the 99th percentile of the distribution that generated the parameter r used in GeoInd. In other words, the user has 99% chance of being remapped somewhere within this distance. For all neighboring points, we compute the posterior and then, we calculate the geometric median of those coordinates using the iterative Weiszfeld's algorithm [52]. The geometric median minimizes the average Euclidean distance and hence, returning the new, optimal (in terms of utility as privacy should remain the same) location.

We implement the Random mechanism tossing a biased coin every time a location is about to be reported. The bias is set so that users

release on average 40%, 60% or 80% of their measurements. For the Release mechanism, we sort all the locations reported by a user in chronological order, and release a new location only if it is separated by (at least) $x \in \{30, 60, 90\}$ meters from the previously reported one. If two locations are less than x meters apart but in different days, we release them both.

Last, we implement Rounding by rounding to 2, 3, or 4 decimals of the latitude and longitude of the users' locations. Effectively, this reduces the location accuracy to roughly 1,100 meters, 110 meters and 11 meters, respectively.

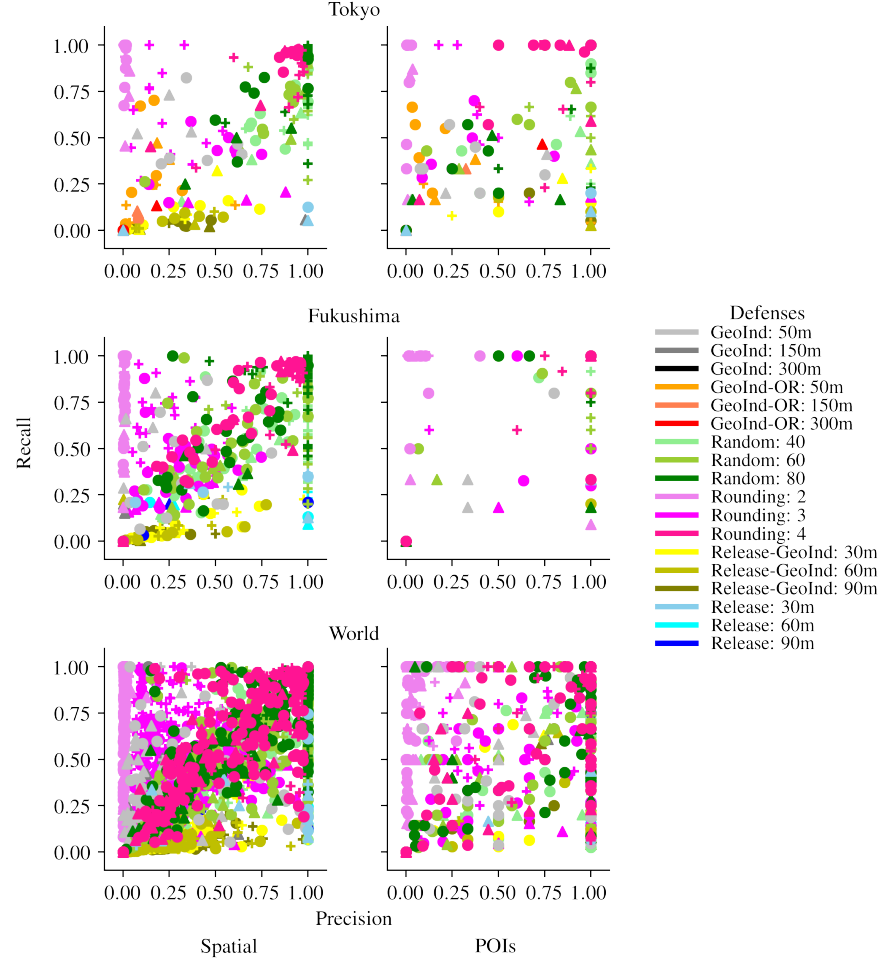


Figure 2.3: Safecast privacy gain: Spatial (left) and POIs (right). Amount of measurements per user + : $<10k$, \bullet : $[10k, 50k]$, \blacktriangle : $>50k$. Each point on the graphs represents one user.

2.6.3 Privacy gain

2.6.3.1 Safecast

We first evaluate the privacy gain of the LPPMs in the Safecast dataset. Figure 2.3 shows the Spatial (top) and POI (bottom) gain for Tokyo, Fukushima, and the whole world. The x-axis represents precision, and the y-axis recall. Each point in the graph represents a user, and the markers' shape indicate the amount of measurements she contributes. The colors represent the LPPMs. Defenses that provide large gains result on points close to the figure axes. Points near the y-axis indicate low precision, i.e., cases in which the adversary correctly identifies some (or even all) of the true locations but also inferred many other wrong locations. Points near the x-axis indicate low recall, i.e., cases in which the adversary correctly identifies some real locations, but misses many others. Unsurprisingly, we observe a high variance in the defenses' performance since it is highly dependent on the user behavior. However, it is possible to identify some trends.

We first discuss the Spatial privacy gain (Figure 2.3, left). For GeoInd we observe that, for the least privacy-preserving parameter ($r = 50\text{m}$), it significantly decreases the number of vulnerable users (grey points in the figure) from the values reported in Table 2.2. The reduction is 50% for Tokyo (from 24 vulnerable users to 12), 45% for Fukushima, and 45% for the whole world. When the mechanism is strengthened ($r = 300\text{m}$), GeoInd adds so much noise (Figure 2.2 for reference) that no users are vulnerable after the defense. In summary, GeoInd seems to provide fairly good privacy gain in Tokyo and Fukushima. Yet, from the whole dataset, it becomes clear that the protection provided by GeoInd is highly dependent on the users' movement patterns. Moreover, we show in Section 2.6.4 that this protection comes with significant utility loss.

The Release-GeoInd (yellow) mechanism works generally better than GeoInd. Even though more users are vulnerable (only between 4% and 13% of the users become not-vulnerable) and the adversary obtains reasonable precision, it yields very low recall. This is because in this method users keep reporting the same obfuscated location until they move. This repetition results in clusters being found on fake locations that often do not overlap with the original ones. This reduction becomes more significant as the defense is configured to provide more privacy (larger z).

GeoInd-OR performs slightly better than vanilla GeoInd. This is because the remapping results on points being repeatedly mapped to popular places causing generation of clusters around those not-real locations. Interestingly, even though this defense was designed to improve utility while keeping the same privacy guarantees, we observe that it decreases utility in the MCS setting (see Section 2.6.4.1).

Similar to vanilla GeoInd, the Release mechanism (blue) significantly reduces the number of vulnerable users – by more than 50% even for the least conservative parameter. However, when precision is very high, i.e., when a cluster is found it corresponds to a real location. The reason is that even though the user hides many point, if a location is visited regularly the user will eventually report enough points around this location to make the cluster identifiable by the adversary.

The Random hiding mechanism (green) does not perform well. First, it reduces the number of vulnerable users less than other defenses (10% decrease in Tokyo, 27% in Fukushima, and only 5% when considering the whole world). From the vulnerable users only a handful obtain good protection. We could not find a clear pattern to predict which movement profiles would best benefit from this defense. For many users, especially those with a few points, removing points at random still yields high precision as the few measurements are very localized. Overall, we do not notice much influence of the fraction of hidden points on the privacy of the users.

Finally, the protection provided by Rounding (pink) depends on the rounding parameter. Keeping 4 decimals reduces accuracy by just 11 meters. Therefore, the adversary finds roughly the same clusters, i.e., for many users we observe high recall and precision after the defense (especially in Tokyo and Fukushima). On the contrary, rounding to 2 or 3 decimals significantly increases the size of inferred spatial areas, which leads to variable recall (depending on the users' movement patterns) and low precision, but at the same time hurts the utility.

Regarding the POI privacy gain (Figure 2.3, right), a first observation is that the amount of users vulnerable to the attack, i.e., points in the graph, is lower. This is because for many users the identified clusters do not contain any POI (according to the OSM API, there could be POIs in reality). Second, for the users who have POIs in their clusters, both recall and precision are higher than in the Spatial gain. This is because many of the large clusters that contribute to the low Spatial precision do not have POIs and thus do not contribute to the confusion of the adversary when identifying particular locations. Furthermore, the clusters that the adversary finds after the LPPMs may cover less area than the original clusters, but still contain most of the users' initial POIs. This provides a higher POI recall than Spatial recall. Third, in this case we observe a significant difference between Tokyo and Fukushima. The reason is twofold. First, the Fukushima prefecture is much larger than the area of Tokyo we consider. Second, Fukushima is a more rural area and thus contains fewer POIs than Tokyo, where even small clusters have many places of interest.

These observations reinforce previous insights that solely considering the spatial dimension [139] may provide a false perception of privacy. Considering a POI-privacy measure is necessary for provid-

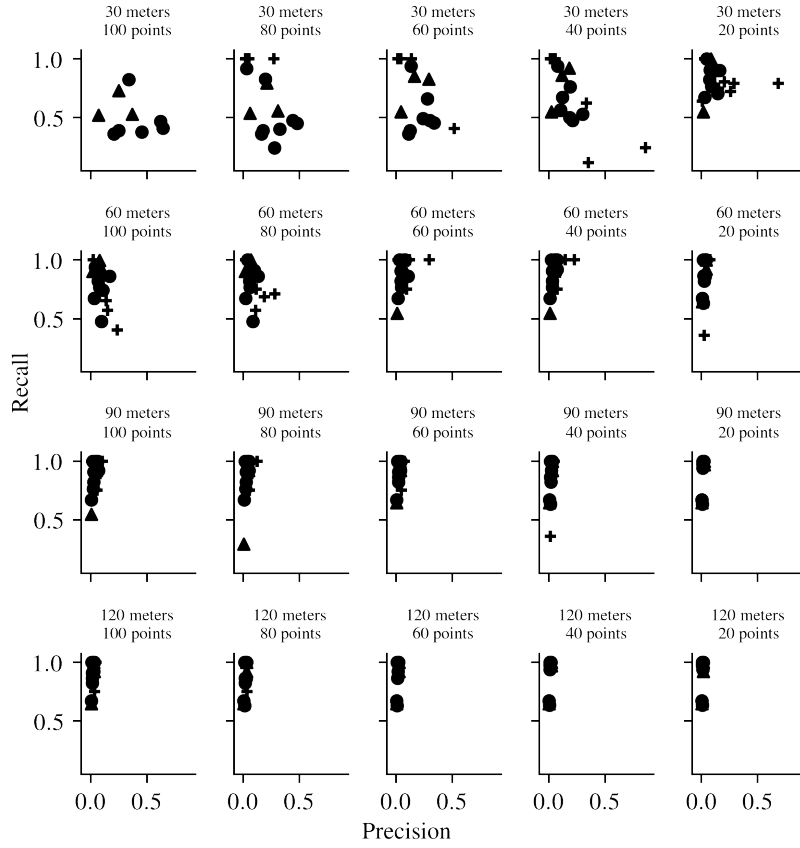


Figure 2.4: Precision and recall for different clustering parameters for GeoInd ($r = 50\text{m}$) in Tokyo. Amount of measurements per user + : $<10\text{k}$, ● : $[10\text{k}, 50\text{k}]$, ▲ : $>50\text{k}$.

ing a comprehensive picture of the privacy threat users face in MCS applications.

Adjusting the clustering parameters. We now study the influence of the DBSCAN clustering parameters on our results. We show the difference in precision and recall for GeoInd ($r=50$ meters) when we vary both the maximum distance and the minimum number of point per cluster in Figure 2.4. As we increase the maximum distance between points and decrease the minimum required points per cluster, the results concentrate on the upper left corner of the diagram. This is because as the parameters become ‘looser’, the resulting clusters grow in size increasing recall (more likelihood of covering all users’ original clusters) but reducing precision due to many false positives. Furthermore, increasing the cluster size increases the adversary’s cost, as the clusters contain a larger number of POIs (Figure 2.5) which requires more filtering and increases the probability of having false positives.

Impact of the amount of measurements on privacy. We present in Figure 2.6 the Spatial gain for the three best LPPMs (all parameters

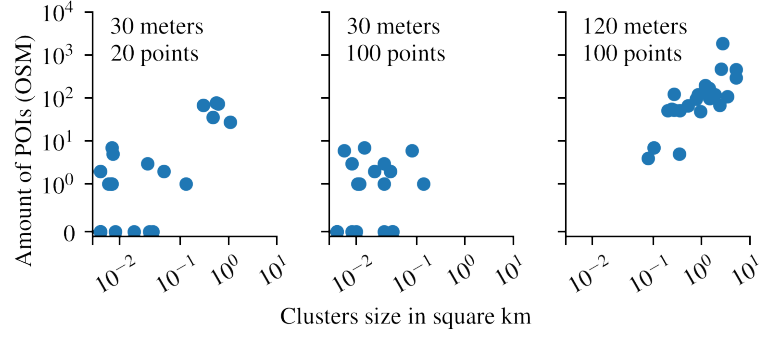


Figure 2.5: Clusters' size and amount of POIs per cluster for different clustering parameters with GeoInd ($r = 50\text{m}$) in Tokyo.

combined) split by the amount of measurements users contributed. We discard Rounding 4 as it does not provide any privacy. We observe that all LPPMs provide low precision regardless of the users' contribution volume. We observe a similar effect for recall except for Rounding which, as explained above, by definition provides variable recall and low precision. In Figure 2.7, the POIs gain (for the three best defenses) validates the results (found in the previous paragraphs) about the differences between spatial and POIs privacy gain. It is important to note though, that for one particular group and defense (GeoInd and users with less than 10K measurements) the results cannot be trusted as they concern only one user.

Counterintuitively, the LPPMs perform worse for users who contribute fewer points. This is because the attack constructs more, and larger (on average 10 times bigger), clusters for people who share many points than for those sharing fewer points. These clusters are split after the LPPMs are put in place, as some reported locations are moved away from their original clusters while other measurements, perturbed with noise, concentrate to new places forming wrong clusters. For Rounding, where every cluster created after the LPPM has roughly the same size, users with a few measurements have higher recall because their initial small clusters are often covered by the large regions resulting from the LPPM.

Thwarting inference attacks. Finally, we evaluate to which extent the different LPPMs are effective at thwarting the inferences reported in Section 2.4, using the *tight* configuration. The LPPMs alter the users' reported locations which in turn may result on deviations in the clusters' centroids locations. We observe that GeoInd and Release-GeoInd fully hide all workplaces, and Release-GeoInd with 30m hides all but one. For Random hiding, we can identify one workplace when releasing 60% of the locations, and 6 when releasing 80%. The Release mechanism is very effective as it hides all workplaces regardless of its parameter. Finally, unsurprisingly, Rounding to 4 decimals does not

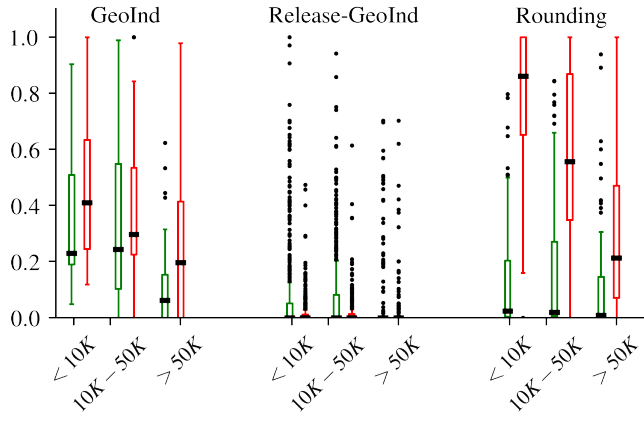


Figure 2.6: Spatial precision (green) and recall (red), depending on the amount of measurements x per user for three selected defenses (all parameters combined), for Safecast.

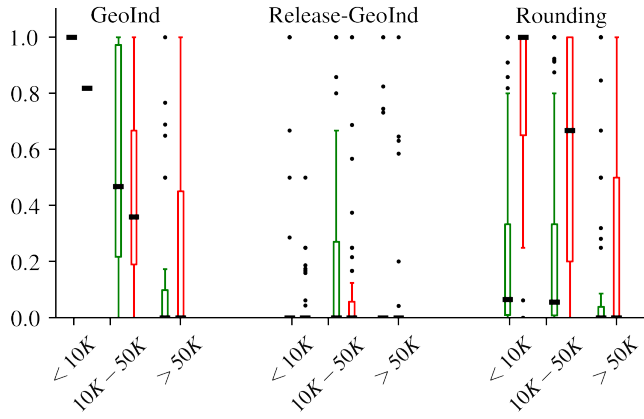


Figure 2.7: POIs precision (green) and recall (red), depending on the amount of measurements x per user for three selected defenses (all parameters combined), for Safecast.

protect against work inference, and Rounding with 3 digits only hides one workplace out of 7. Rounding 2 protects all users.

Summarizing the Safecast privacy results, it becomes clear that some defenses do not work while others, such as GeoInd and Rounding can offer adequate protection with properly tuned parameters. However, as it is discussed in the forthcoming sections, one has to take the utility loss into account, when deciding which parameters to use.

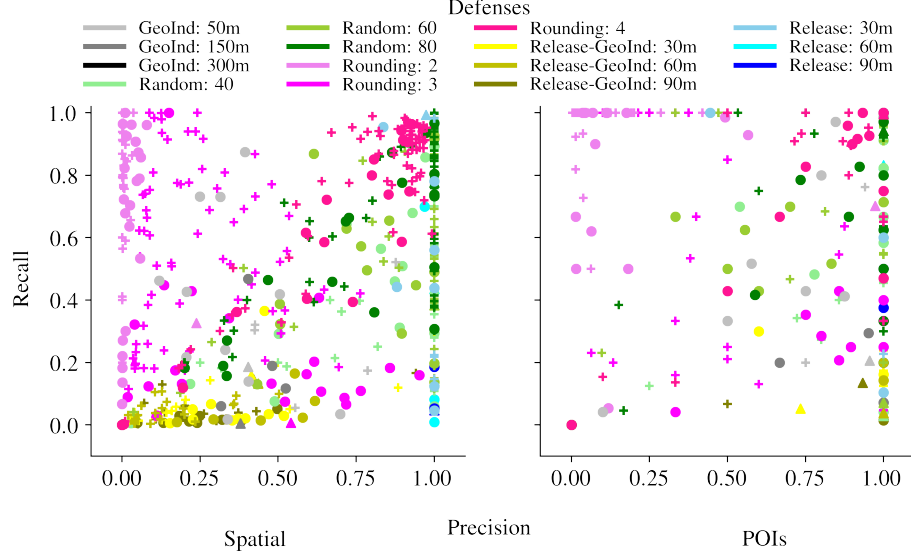


Figure 2.8: Spatial privacy gain (left part) and POI privacy gain (right part) in Radiocells. Amount of measurements per user + : <10k, • : [10k,50k], ▲ : >50k. Each point on the graphs represents one user.

2.6.3.2 Radiocells

Users in Radiocells have on average fewer measurements than those in Safecast. Thus, clustering requiring 100 points yields very few clusters. Hence, for this dataset we loosened the requirement to 25 points per cluster.

We see in Figure 2.8 that, in terms of Spatial gain, GeoInd-based mechanisms behave similarly to the Safecast case. Vanilla GeoInd decreases the number of vulnerable users by 14%, and Release-GeoInd by 2%. Given that only 16% of the users are initially vulnerable this reduction is significant. For the users for which the algorithm finds clusters, the behaviour is the same: GeoInd provides highly variable protection, and Release-GeoInd yields low recall while precision depends on the user behavior. For the hiding mechanisms, the Random and the Release mechanisms decrease the number of vulnerable users by 7% and 14%, respectively. For the vulnerable users, contrary to Safecast, these mechanisms consistently yield high precision, i.e., they offer poor privacy protection for Radiocell's users movement profiles. Finally, the Rounding mechanisms with parameters 2 and 3 offer rea-

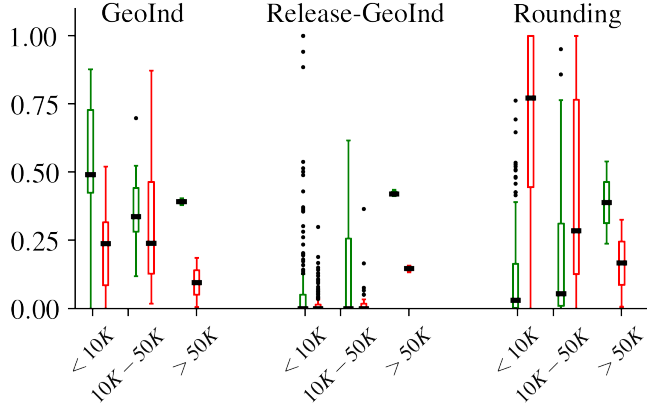


Figure 2.9: Spatial precision (green) and recall (red), depending on the amount of measurements x per user for three selected defenses (all parameters combined), for Radiocells.

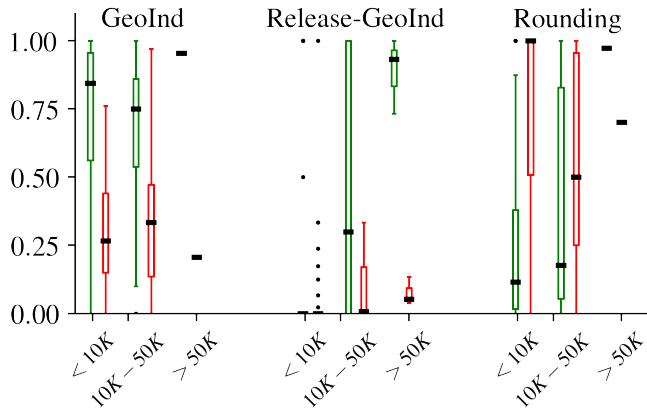


Figure 2.10: POIs precision (green) and recall (red), depending on the amount of measurements x per user for three selected defenses (all parameters combined), for Radiocells.

sonable privacy. Regarding POIs, we observe similar behavior than with the Safecast dataset. For reference, in Figures 2.9 and 2.10 we represent the impact of measurements on the privacy of the users. The results confirm the observations made for the Safecast dataset with the exception that this dataset has less users and they have on average much less measurements. Hence, for some groups of contributed data the results are generalized by a small amount of users.

Overall, the results in Radiocells are consistent with our findings in the Safecast dataset, confirming the trends regarding the LPPMs' behavior in the MCS setting.

2.6.4 Utility-privacy trade-off

In order to accurately evaluate the effectiveness of LPPMs one also needs to pay attention to the incurred utility loss, as this is a crucial factor for the successful adoption of any privacy-preserving methodol-

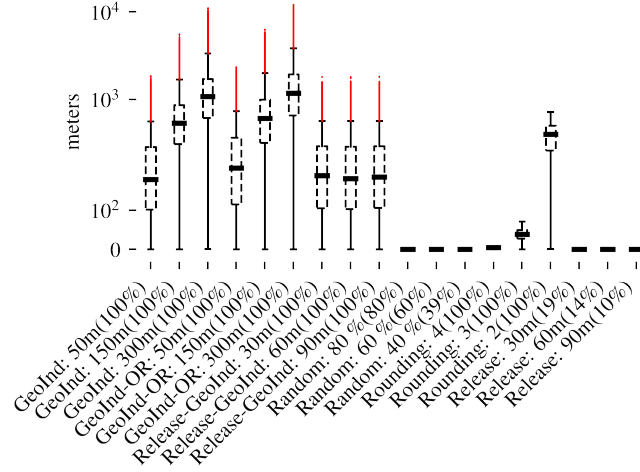


Figure 2.11: Measurement error in Tokyo using a distance-based metric. This can be interpreted either as privacy gain or utility loss.

ogy. As already discussed though, in MCS the utility loss is different than the more traditional LBSs setting.

2.6.4.1 Safecast

We now study the utility loss for Safecast.

Distance-based metric vs aggregate statistics for MCS. We first evaluate the utility loss incurred by the LPPMs measured using the LBSs-oriented distance-based metric described in Section 2.5.3. Figure 2.11 presents the per-reported location utility loss for users in the Safecast-Tokyo dataset. The y-axis indicates the distance in meters, and the x-axis the LPPM and the percentage of points that are released. This utility metric is based on the distance between reported and real location, but disregards the (radiation) values that Safecast cares about. Random and Release LPPMs seem to be the best, and GeoInd LPPMs offer the worst performance as they tend to spread measurements, sometimes more than a kilometer away from the initial measurements (see Figure 2.2 for reference)

Figures 2.12 and 2.13 show the per grid-point utility loss for Tokyo and Fukushima, respectively, when the utility loss measured as the difference between radiation values on the generated map. We observe a similar behavior in both regions, though the median loss in Fukushima is slightly higher and it has many more outliers (reaching up to 10^4 difference) than in Tokyo. In this case, after the interpolation all GeoInd variants offer roughly the same utility loss on average. However, Hiding and Rounding strategies offer better performance, yielding smaller errors for the least protective parameters.

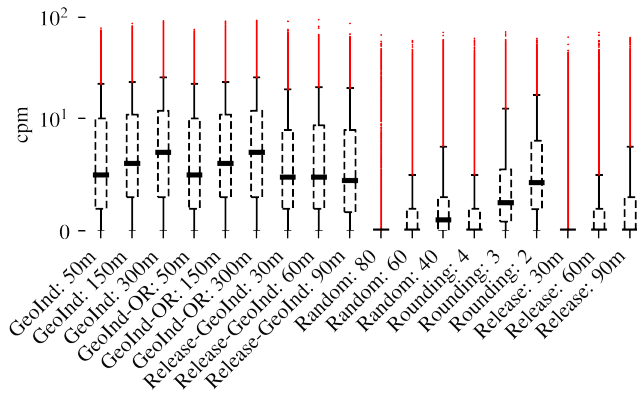


Figure 2.12: Absolute difference in Tokyo's radiation values with Safecast dataset.

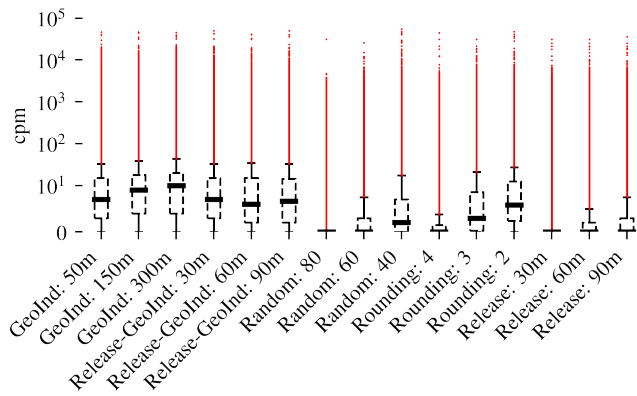


Figure 2.13: Absolute difference in Fukushima's radiation values with the Safecast dataset.

If we compare the distance-based results (Figure 2.11) to the utility loss computed considering the values to appear in Safecast’s interactive map (Figure 2.12) we find significant differences. A first observation is that, LPPMs based on GeoInd fare the worst in the distance-bases measure compared to the aggregate statistics. Furthermore, distance-based metrics do not properly quantify the loss of LPPMs that hide some points (Random and Release). While it is true that the released points have no error, the many points that do not appear incur in a loss. This is made evident by the aggregated metrics, where we observe that the more points are hidden, the larger is the utility loss. We note that relying on Markov mobility models such as in [76, 140] could help interpolating the hidden locations. Yet, this would not help recovering the (radiation) values attached to them and the utility loss could not be avoided. For non-hiding mechanisms, distance-base metrics consistently result on larger median loss, though less variance and less outliers.

In summary, distance-based metrics provide a very different perception of the LPPM performance than considering utility functions computed on the geo-located values, overestimating the performance of some methods (e.g., hiding strategies) and underestimating others (e.g., GeoInd-based LPPMs). We conclude that traditional LBSs-oriented metrics are inadequate for measuring utility in MCS scenarios.

Why optimal remapping does not work for MCS. Even though GeoInd-OR was designed to increase utility while preserving privacy, we observe that in the MCS case utility roughly stays the same (Figure 2.12), and privacy slightly increases, both in decreasing the number of vulnerable users and in increasing the spatial gain. The reason for this mismatch is that this mechanism was designed in the context of LBSs, where remapping locations to places where the user is likely to be is bound to provide good utility on average. However, in Safecast the utility does not depend on the locations themselves, but on the associated measurements. Remapping the location, however, concentrates measurements in these popular locations effectively polluting the measurements. To illustrate this effect, we show in Figure 2.14 a heatmap of the prior probability of users’ locations over all locations in Tokyo. This figure was obtained by rounding every reported location within the prefecture of Tokyo to 3 decimal digits (for visualization purposes) and then counting the occurrence of each unique location. In this way, we have a visual representation of which places attract the most users and which are rarely visited. We observe that the probability is low (represented in white) for most places with just a few exceptions, which are popular locations for users who contributed lots of measurements (represented in black). Remapping in the low probability areas has a randomizing effect, since most locations have the same probability. On the contrary, when remapping happens in a region with a location with high probability, this popular location con-

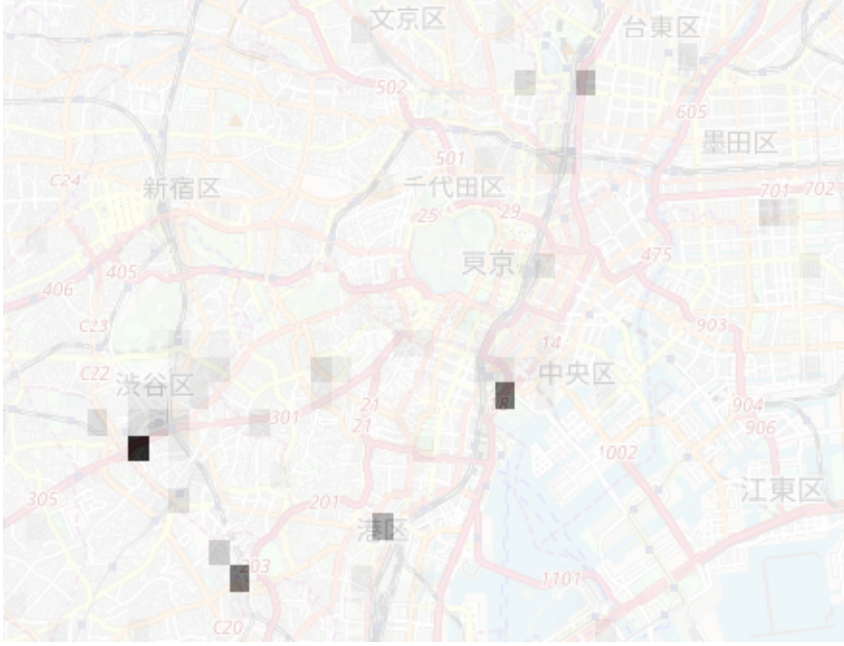


Figure 2.14: Prior probability of visiting locations in Tokyo (white - low probability, black - high probability).

centrates the remapping. This, while hurting utility increases privacy, as it creates artificial clusters that reduce the adversary's precision and recall.

Semantic interpretation. The absolute difference in cpm of measurements before and after the defense gives a rough idea about the utility loss, but they are difficult to interpret. Are they significant? What is the effect of outliers? Does reporting the values after the defense have any implication on the danger for human health? To answer these questions, we study how the variance introduced by the defenses can change the interpretation of the risk at a given location. To this end, we rely on the cpm safety scale [27] provided with one of the top-seller Geiger counters (radiation measurement devices) on the market. This scale contains five categories:

- Category 1: 0-50 cpm. Normal radiation background.
- Category 2: 51-99 cpm. Medium level.
- Category 3: >100 cpm. High level.
- Category 4: >1000 cpm. Very high level, leave area.
- Category 5: >2000 cpm. Extremely high level, immediate evacuation.

We select the prefecture of Fukushima and two defenses that produce a good level of privacy: GeoInd 300m and Rounding 2. For each of the 2.25 million grid-points on Safecast's radiation map for Fukushima, we compute their radiation category according to the safety scale before and after each defense. For GeoInd 300m, which is of probabilistic

Table 2.3: Category changes, for each of the radiation danger groups, after applying Geo-Ind ($r = 300$ meters) in Fukushima.

Geo-Ind: 300	1	2	3	4	5	Number
Original						of points
1	79.7%	19.3%	1%	0.003%	0.001%	1354110
2	41.5%	49.5%	9%	0.023%	0.01%	650486
3	8.7%	35.9%	52.2%	2.3%	0.9%	229848
4	2.5%	3.3%	49.3%	29.8%	15.1%	10489
5	3.9%	1.7%	34.7%	29.3%	30.4%	5067

Table 2.4: Category changes, for each of the radiation danger groups, after applying the Rounding mechanism (2 decimals) in Fukushima.

Rounding: 2	1	2	3	4	5	Number
Original						of points
1	89.3%	10.3%	0.3%	-	0.001%	1354110
2	30.2%	64%	5.8%	0.003%	-	650486
3	0.7%	22.6%	74.8%	1.6%	0.3%	229847
4	0.2%	0.01%	43.3%	39.6%	16.9%	10490
5	0.9%	-	9.3%	42.1%	47.6%	5067

nature, we repeat the procedure 10 times and report the average. We present the results in Tables 2.3 and 2.4.

We observe that the majority of the points either stay in their original category or move to a nearby. However, we observe some extreme category jumps from the first category (safe radiation levels) to the fourth and fifth (high danger). For instance, GeoInd causes 53 places to be marked as dangerous instead of safe. Even more alarming, 283 locations that should be marked as extremely dangerous are marked as safe or slightly elevated (categories 1 and 2). On the contrary, the Rounding mechanism limits the number of extreme changes. For instance, there is a category jump from 5 to 1 and 2 only for 45 grid-points.

The case of high precision measurements. One of the reasons why Safecast collects crowdsourced measurements is the monitoring of radiation “hotspots” that could be dangerous for public health. When locating hotspots, precision is important both to understand their implications and to keep low costs if experts have to be sent to study the origin of the abnormality.

To understand the impact of LPPMs on hotspot location we perform detection in Tokyo and Fukushima by looking for locations with more than 100 cpm radiation after averaging the measurements over the last

270 days but *before interpolating the data*. This is to avoid that interpolation modifies the position of the hotspots, or even eliminates them. We show the results of detection when using the raw measurements, and after the application of several defenses in Figures 2.15 and 2.16.

We see that noise-based mechanisms spread the measurements and create additional hotspots and, as the noise increases, so do the hotspots. This renders the hotspot detection useless for Safecast as the results cannot be properly interpreted. For instance some hotspots could originally be in places known to present high radiation, and hence be already closely monitored by the authorities. The spreading of the points conveys a much different message, especially when the hotspots appear in zones considered to have low radiation in the past.

Generalization such as Rounding 2, which provides a good privacy-utility tradeoff for aggregated statistics, also performs poorly. In this case the defense causes hotspots to disappear, potentially causing a dangerous situation if a high radiation zone is marked as safe. We also carry out experiments with Hiding mechanisms and we find that, similarly to Rounding, they miss some of the original hotspots.

Takeaways. Considering only the privacy loss, see Section 2.6, GeoInd variants (except GeoInd 50m) and Rounding to 2 decimals seem to offer the best performance, while Random sampling and Release’s provided protection is too dependent on users’ movement profiles. However, an analysis of the utility impact, in particular the semantic interpretation or the case of high precision measurements, indicates that *none of the existent LPPMs is well suited* for the Safecast setting. The semantic interpretation results indicated that even if two defenses produce similar average results, the outliers they create can convey opposite messages. Furthermore, even a slight addition of noise or generalization can hinder the project’s ability to correctly locate abnormal events. These limitations effectively prevent Safecast from deploying them and protect their users’ privacy.¹¹

2.6.4.2 Radiocells

Radiocells’ utility function is rather different than the one for Safecast. Instead of averaging measurements associated to a location, Radiocells averages all reported coordinates associated to an antenna to locate its position. We show the utility loss for different LPPMs when performing this task in Figure 2.17.

All GeoInd variants result in high utility loss, with medians between 80 and 400 meters, and with outliers beyond 2 kilometers. Surprisingly, in this use case hiding mechanisms (Release and Random) have many outliers. After manual inspection, we found out that several Radiocell users had inconsistent measurements. For instance, a user was swapping her measurements’ longitudes and latitudes in a random

¹¹ This statement was verified in communication with Safecast.

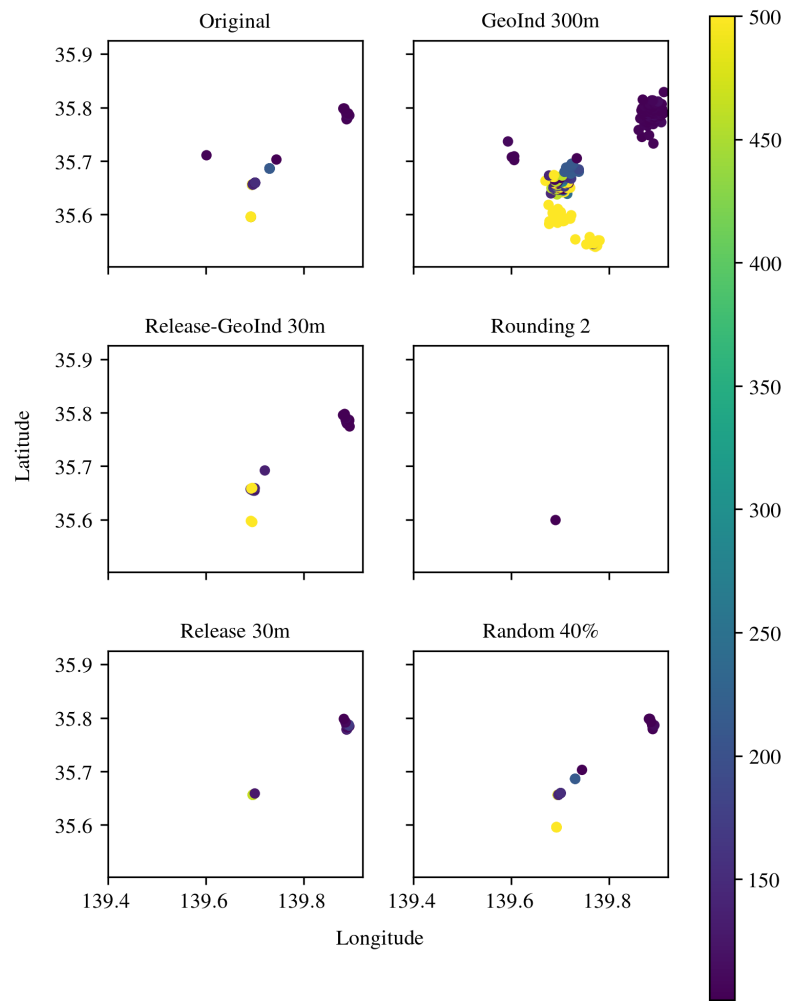


Figure 2.15: Safecast: Hotspot detection for areas in Tokyo with at least 100 cpm. Comparison of various defenses vs the original hotspots.

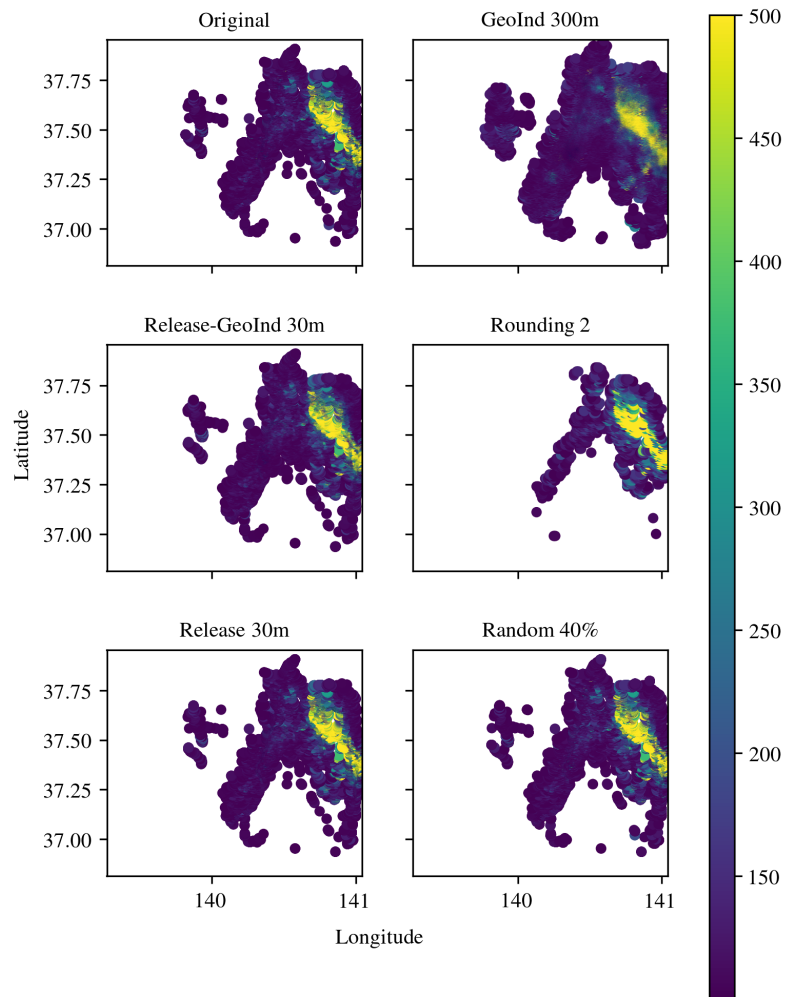


Figure 2.16: Safecast: Hotspot detection for areas in Fukushima with at least 100 cpm. Comparison of various defenses vs the original hotspots.

pattern. Other outliers are caused by providers moving their antennas creating mixed measurements for a given ID. For these cases, hiding LPPMs may greatly affects the average computation. Furthermore, hiding defenses may also influence the number of antennas located. In our dataset we detect from 10.2% up to 18.6% fewer antennas when the Release defense is used, and the Random mechanism eliminates from 2.6% up to 13.7% of them.

The best mechanism in the Radiocells dataset is Release GeoInd which offer on average lower utility loss than other LPPMs and provide acceptable privacy. However, some antennas might be moved over a kilometer away. The next best alternative is Rounding 2, that has a higher median utility loss, but no outliers. However, as the goal of the project is to *accurately* detect antennas in order to give individuals the ability to geolocate themselves offline or to enable scientific studies, a median error of 100 meters (Release GeoInd) or 200 meters (Rounding 2) is considered too large and precludes Radiocells from deploying them.

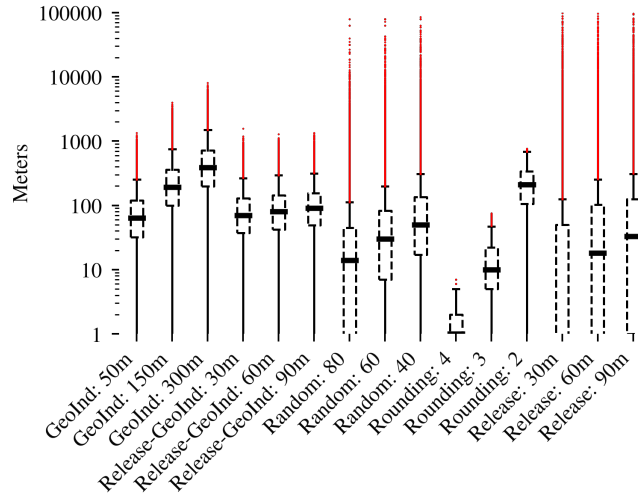


Figure 2.17: Radiocells: Utility loss (distance from original tower location).

2.7 TOWARDS PRIVACY-PRESERVING MOBILE CROWDSOURCING

As our goal is to empower MCS application developers for more privacy aware system designs, we illustrate how parts of this chapter can be used as a tool for privacy analysis of location data.¹² Moreover, we elaborate on technical and non-technical steps to enhance privacy at smaller utility cost in the context of MCS applications.

¹² The code for the evaluation is available at <https://github.com/sboukoros/Location-Privacy-Framework>

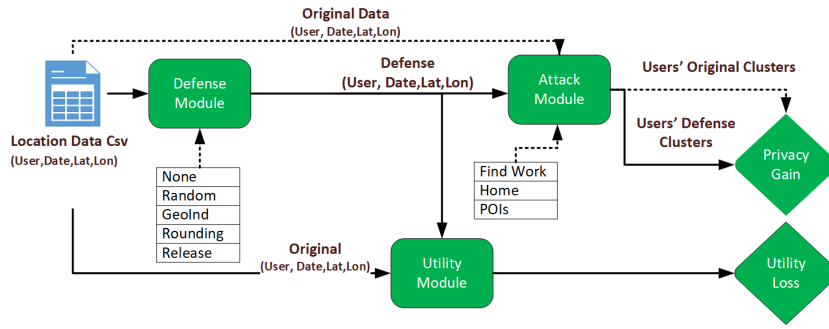


Figure 2.18: Systematic privacy evaluation of mobile crowdsourcing. Green color denotes the modules that can be updated.

2.7.1 A tool for the systematic evaluation of LPPMs for MCS

To conduct the study described in this chapter, we implemented three different modules: one for the attacks (Section 2.4), one for the defenses (Section 2.5), and one to evaluate the defenses' privacy gain (Section 2.5.2) and utility loss (Section 2.5.3). These modules can be combined to build a tool (see Figure 2.18) to enable application developers (as well as users) to understand the privacy risks stemming from sharing geo-located data with crowdsourcing applications. The tool can then be used to evaluate the effectiveness of existing privacy-preserving mechanisms and take an informed decision depending on their impact on the application's utility.

The tool works as follows. It receives as input the original measurements reported to the crowdsourcing application as a Comma Separated Values (CSV) file, where each line represents a geo-located value. In our current implementation the file contains the following fields: User ID, Captured Time, Latitude, Longitude, offset from UTC time and Value. This file is input to the defenses module where a selected defense, or None if one wants to evaluate the privacy risk associated to the publication of the raw data, is selected to modify the location of the entries. Note that, for certain defenses, this entails eliminating rows. Both the original and the defense file(s) are input to the utility function module that is instantiated for the particular application under study, and to the attack module. The outputs of these modules are the utility, and the clusters for the raw and modified data. Using any available API (in our case OpenStreetMap Overpass API) the tool retrieves the POIs inside these regions. With this information the tool can compute the privacy gain and the utility loss and display them in a way understandable by the users (e.g., with plots).

All modules in the code are implemented as standalone functions or scripts. This way, new attacks, defenses, or metrics can be easily added to our framework by researchers and developers. Also, the modules can be integrated into mobile crowdsourcing applications design frameworks to support privacy practices in software engineering.

2.7.2 *Towards effective defenses*

In this section, we discuss possible strategies to improve the trade-off between users' privacy and MCS utility.

An unexplored approach is the use of advanced cryptographic protocols to compute the values of interest for MCS without revealing the users' individual values to the providers [59]. For instance, users could use multi-party computation to collaboratively compute aggregates and only report the result to the provider. However, cryptographic approaches require high computational power on the users' side and increase the bandwidth needs to perform the joint computation. Furthermore, this would limit the availability of raw measurements for analysis others than those predefined by the cryptographic protocols, which is at odds with the principles of open data and open science defended by most of the MCS platforms.

In our evaluation, we only considered spatial generalization. Another avenue to explore would be to also generalize the time dimension. While on its own time obfuscation cannot hide patterns revealed by repeated visits, this technique could be combined with full de-identification and hiding of users to mitigate the inferences of the adversary. For instance, the MCS service provider could release a batch of measurements once a day or once a week without linking these to any user identifier. These techniques would be cheaper than the use of cryptography, but require to trust the service provider to apply the sanitization.

A third research path is the co-design of defenses and aggregation algorithms. In this chapter, we have considered that the output of the LPPMs is directly input to the utility functions currently used by MCS providers. However, it would be possible that the providers adapt their data processing to account for noise, using statistical methods or machine learning, as done in fields that rely on noisy sensors [137] or train in different settings from which they are deployed [62, 121, 149].

Finally, MCS could provide users with dedicated local software (e.g., building on our evaluation method) to alert them regarding the privacy dangers of publishing raw location data. Such a system would allow them to selectively hide some of their measurements, reducing the confidence of inference attacks. We note that, when building such a tool, one would like to consider attacks beyond the POI-based inferences, such as co-location attacks [120].

2.7.3 *Best practices for mobile crowdsourcing developers*

During this study, we identified a number of issues related to the collection and sharing of data that, even though cannot fully prevent inference, could make inference attacks detectable and could render potential attackers accountable.

A first consideration to make is the type of policy under which MCS publish the collected data. While making large datasets available to everyone for unrestricted use is admirable, and certainly of high value for the academic community, it can have serious implications for the altruistic contributors. To reduce this risk, developers could add clauses to the policies that not only mandate that use of the data is properly acknowledged, but also that it is well documented, implying that researchers or other individuals have to disclose how they have processed the data and for which purpose.

Second, both Safecast and Radiocells datasets are available for download without the need for authentication. This hinders traceability of who has the data, and thus enables stealthy attacks where neither the users nor the applications are aware of the danger. Like in other projects that make data available for research and other purposes (e.g., the Drebin project¹³), these sites could require simple registration to maintain a log of who has had access to the datasets. Together with the previous requirement, which would include documentation of sharing, it should help mitigating the risks.

Third, these applications typically do not perform any control on who are the contributors. This poses a particular problem when it comes to children. In many jurisdictions, children's data are subject to particular legislation [134, 150], and in particular require the parents' consent to be collected and processed. The lack of controls upon collection implies that the datasets could contain children's geo-located data collected illegally. Adding controls would solve this problem and also support the previous two points.

Finally, the datasets we studied contain data from users from all over the world. These users, therefore, are subject to different legislations that regulate how their data can be processed. While this may not be a problem for corporations or criminals that want to exploit the datasets, it creates a hurdle for researchers who have to obtain approval from their institution for data processing. Lack of proper documentation may limit the free use of the data for science, effectively hindering one of the main goals of these applications. Better documentation as to the origin of data and its use possibilities would greatly facilitate the process.

2.8 CHAPTER SUMMARY

In this chapter we explored the applicability of location privacy LPPMs in the context of mobile crowdsourcing. For our experimentation we used data and utility functions from actual crowdsourcing projects. We first validated the privacy danger of publishing raw location data by performing inference attacks on the Safecast dataset. Then, our results in the quantification of privacy gain and privacy-utility trade-off,

¹³ <https://www.sec.cs.tu-bs.de/~danarp/drebin/>

illustrated that traditional LPPMs are not applicable in the scenario under study as they have been optimized for a different use case. Furthermore, we provided guidelines and showed how our work can be compiled into a tool to enable further research into privacy preserving mobile crowdsourcing and guidelines to application developers of MCS projects.

PRIVACY ASSESSMENT OF AGGREGATION SCHEMES IN SMART METERING

Published Content

N. Büscher, S. Boukoros, S. Bauregger, and S. Katzenbeisser.
 "Two is not enough: Privacy assessment of aggregation schemes in smart metering" In: Privacy Enhancing Technologies Symposium, 2017, pp. 198–214.

3.1 SMART METERING

Smart meters are the next generation electricity measurement devices which also have networking capabilities. They allow energy suppliers a permanent and detailed monitoring of their customers' energy consumption in order to reduce costs through more efficient and automatized power management. Besides the obvious advantages for energy suppliers, the expected increase in renewable energy, electric cars and prosumers (consumers that also produce) in the following decades, require more reliable and flexible energy networks. Such networks consisting of smart meters are called smart grids. A smart grid is basically an advanced, multi-directional power network and is regarded as the future of energy supply systems. In recent years, many countries worldwide introduced laws in order to expedite the use of smart meters in households. An example is the EU Directive 2006/32/EC, which asks all EU member states to provide "individual meters that accurately reflect the final customer's actual energy consumption and that provide information on actual time of use" for energy consumers [123].

Despite the economical and ecological advantages for the involved parties, the widespread information flow from energy consumers to producers is a serious threat to the consumers' privacy. The establishment of smart meters generates sensitive data to an extent that could not be reached using conventional meters. The continuous disclosure of energy consumption data in conjunction with algorithms like non-intrusive appliance load monitoring (NILM) [85, 100], helps third parties to figure out daily routines of households, particular appliance uses, individuals' presence in a building or even the movie playing on the television [82]. If marketing agencies collude with energy suppliers, they can gather detailed information regarding household

appliances [101]. It's not hard to imagine that these data can be used for targeted advertisement campaigns, new offers, etc. Criminals who are able to tap into a meter's data management system could predict when the occupants of a building will not be present [102]. Therefore, they can orchestrate their illegal activities more accurately. Even worse, mass surveillance is significantly enhanced. With little resources, interested malicious parties can observe the daily routines of millions of households.

These privacy concerns have been known to academia, industry and governmental institutions for years and therefore, a plethora of privacy mechanisms have been proposed to protect the consumers' privacy in the smart grid. The most promising and well researched privacy mechanisms are based on aggregation schemes, e.g., [78, 99, 103, 107, 111]. The core idea is to form groups of devices within the smart grid. Then, only the aggregated power consumption of the group is periodically reported to the energy supplier. The aggregate of a group can securely be computed using either a trusted third party, or preferably through cryptographic means, e.g., partial homomorphic encryption, secret sharing or other secure computation techniques. This solution has also been suggested as the-way-to-go by an expert group, set up by the European Commission [159].

Even though secure aggregation is technically solved, a major question has, to the best of our knowledge, barely been addressed. Namely, which aggregation size (number of smart meters in every group) is required to achieve privacy for consumers. During the smart meter roll-out in the United Kingdom, a study conducted by the industrial body "Energy Networks Association" concluded that aggregating the consumption of *only two smart meters* provides sufficient customer privacy [116]. However, this result seems to be elusive. It is not hard to imagine two households, where one person works during day shifts, while the other during night shifts. An aggregate of the two load profiles is protecting neither household because the two individuals will most likely be at home and use their appliances at different times.

For the aforementioned reasons, in this chapter we evaluate the privacy achieved by smart meter's aggregation schemes. Such a study, benefits users who soon many have to install smart meters in their home as they will be able to understand the privacy risks involved even in the case of aggregate reports. Additionally, it sheds light to the dilemma of whether two (or just a few) smart meters in aggregates offer adequate privacy.

Chapter outline

The rest of this chapter is organized as follows. We discuss related work in Section 3.2, before introducing our metric in Section 3.3. Then, in Section 3.4 the analyzed datasets and the evaluation approach are

described. In Section 3.5 we apply the proposed metric on real energy consumption datasets and present various case studies. Furthermore, in Section 3.6 we study the diversity of energy consumption and the applicability of generated load profiles for privacy research.

3.2 RELATED WORK

3.2.1 *Privacy mechanisms for smart metering*

Various privacy enhancing technologies have been proposed for smart metering. Fhom et al. [71] proposed the use of a third party that anonymizes the data and empowers users. This party is instantiated as a software solution inside the users' premises. Users can specify their privacy requirements, get feedback regarding their privacy levels based on the data they share with the energy providers, and specify by which entities their data can be processed. Kalogridis et al. [91] proposed to blur the load signature of individual smart meters – that is the unique patterns of every load profile, in order to achieve privacy. This can be achieved by the use of a battery and an algorithm that distributes energy accordingly.

Cryptographic solutions have also been proposed for smart metering. Efthymiou and Kalogridis [67] proposed a third party escrow architecture, that anonymizes consumption data by assigning two different IDs to the data; an anonymous high-frequency ID and an attributable low-frequency ID. The latter is assigned to low frequent consumption data needed for billing, while the high frequent data used for continuous demand monitoring is attributed with the HFID. Molina-Markham et al. [114], using actual smart meters load profiles from three houses demonstrate how without any prior knowledge an adversary can infer basic information such residents' presence, the amount of people in the building, sleep patterns or single appliance's usage. Then, using zero-knowledge proofs, they proposed a privacy preserving protocol for smart metering billing. Using the proposed protocol the provider can only learn the final cost of the electricity consumed but not the power profile. However, as it is necessary for the provider to have some knowledge of load profiles for analysis and prediction, the protocol sends high resolution data to the provider in aggregates. Chim et al. [49] proposed a system for privacy-preserving authentication with tamper resistant devices and pseudo-identities. Through homomorphic encryption, the system also supports smart meter reports' aggregation within regions. Their system assumes however, that the energy provider is a trusted party which knows the actual identities of the users and hence, their energy consumption. Danezis et al. [55] designed privacy-preserving protocols for a variety of smart metering operations. Their models, based on multiparty

computation with secret sharing, demonstrate how operations such as fraud detection or advanced statistics are feasible.

Differential privacy mechanisms have been explored in the context of smart meters. Acs and Castelluccia [30] proposed a distributed noise addition mechanism in order to avoid third parties. In their scheme smart meters are clustered to groups and each smart meters transmits its data encrypted after it has added noise. In such a way the provider cannot obtain values from independent smart meters. Backes and Meiser [35] also explored the use of rechargeable batteries as means to provide privacy however, with differentially private guarantees. Their system adds or subtracts noise to the electrical grid monitored by the smart meter, while recharging the battery. However, the latter happens in a way that cannot be inferred by the smart meters logs as if the usage pattern of such a battery is inferred, it can be detrimental to privacy. Zhang et al. [158] also proposed a battery-based differentially private scheme. However, in their work they aimed to maximize cost saving under static and dynamic pricing policies.

Aggregation of smart meters reports is a well explored privacy mechanism. Usually, the aggregation happens either by a trusted third party or better by cryptographic protocols. Garcia and Barts [78] proposed an additively homomorphic encryption scheme for detecting fraud in smart metering. They rely on a concentrator device that provides energy to a cluster of smart meters and wants to detect inconsistencies between the aggregate report of the smart meters and the energy it has actually provided. In order to check for such energy leakages, the smart meters via homomorphic encryption and the help of the concentrator, securely aggregate their results in regular intervals and report the collective consumption. Then, with a simple comparison between the reported value and the energy load the concentrator had provided to the meters, leakage is detected. Marmol et al. [111] proposed an aggregation scheme based on bihomomorphic encryption, i.e. an encryption scheme additively homomorphic in plaintext and key space. In their mechanism, every smart meters encrypts and transmits their data over a secure channel. The provider, whose key is an aggregate of all the smart meters' keys, can only decrypt the data after aggregating them. Li et al. [103] also rely on homomorphic encryption to aggregate smart meters reports. The reports are incrementally aggregated and encrypted over paths that cover all the smart meters until they reach the energy provider. Lu et al. [107] proposed a homomorphic encryption scheme with better performance than previous attempts. Their scheme is based on optimizations in terms of communications and allows aggregation of multidimensional data. In the same direction, Kursawe et al. [99] presented four different aggregation based privacy mechanisms using various cryptographic approaches achieving lower complexity than previous mechanisms.

3.2.2 Privacy quantification

Various works aim to quantify privacy in the context of smart grids. Lisovich and Wicker [106] investigated privacy in smart metering and more specifically, what an adversary can infer about residents' habits including, presence or absence, sleep and wake cycles and appliance usage. They processed load profiles with behavior extraction algorithms and compared their results with those of actual cameras. They quantified the privacy loss using metrics about correct inferences, i.e. whenever the camera was detecting something and the results of the algorithm would agree. Their results portray that basic routines can be inferred with high accuracy.

Dong et al. [60] quantified the utility-privacy trade off in smart metering. Their scheme works under the assumption that users' energy consumption follows a hierarchical Bayesian model where users have some private behavior, this behavior is reflected on the users' device usage patterns and hence affect the final load profile. They introduced a privacy metric called inferential privacy that in essence, measures the adversarial error where the adversary tries to infer users' private behaviors (such as usage of specific devices), having access to the smart meter's current and past reports, and knowledge of how various types of consumers use their devices. Eibl and Engel [68] examined the effect of differential privacy on the utility of smart meters' aggregated reports. Using actual user data, they added laplacian noise to aggregates and studied the effects of different differential privacy parameters and the number of smart meters in the aggregate. Their results identify several open issues in the application of differential privacy such as the determination of the correct privacy parameters and the amplitude of the noise in small aggregates. Shankar et al. [136] developed a formal framework for smart metering where they evaluated privacy and utility. In their work, they modeled smart meters' measurements using hidden markov models. Their utility metric is based on the distortion function between real and perturbed data. They measure privacy loss via the mutual information between the original time series and the one after defenses are in place.

Our contribution

The privacy modeling proposed in this chapter is based on the cryptographic game developed by Bohli et al. [42]. The goal of their game is the evaluation of privacy protection mechanisms for a group of smart meters. The privacy level provided by the smart meter application is defined as the advantage of an adversary over random guessing, when distinguishing two groups of smart meters and their protected load profiles. In contrast, this chapter explores the creation of a cryptographic game to *isolate individuals in aggregation schemes* of various

loads. In our work, we use a more realistic adversarial model and real world datasets. Moreover, using our proposed game we further explore various properties of the aggregates and their impact to privacy.

3.3 AGGREGATION PRIVACY MODEL

As discussed, many cryptographic schemes have been proposed that allow the privacy preserving (provably secure) computation of smart meter aggregates. However, only a few metrics have been proposed that assess the effectiveness of smart grid privacy protection mechanisms in a formal and sound manner.

We propose such a framework, but first we formalize data aggregation in the smart grid before we iteratively develop our privacy metric.

3.3.1 *Smart grid aggregation model*

We make use of the following abstraction, which models the interaction between smart meters and an energy supplier. Informally speaking, when using privacy-preserving aggregation schemes, the energy supplier should learn the aggregated power consumption of groups of smart meters in every measurement period. For simplicity, we reduce our model to a single group of meters. Thus, the model consists of an energy supplier ES and a group (set) of smart meters $S = \{s_1, s_2, \dots, s_n\}$ with $n > 1$. For practicality, we further assume a virtual party, the aggregator V, which connects all smart meters in S with the ES. In practice, this aggregator can either be instantiated by a trusted third party or by a cryptographic aggregation protocol, established between the smart meters. Moreover, a discrete notion of time $T = \{1, 2, 3, \dots\}$ is used. In each time period $t \in T$, every smart meter s_i is attributed with a power consumption value $e_{i,t} \in \mathbb{R}$, where \mathbb{R} is the set of possible readings from a power consumption meter. Furthermore, we refer to consecutive consumption values as load profile. We denote a load profile of length l for a single smart meter s_i with $\hat{e}_i(l) = (e_{i,1}, e_{i,2}, \dots, e_{i,l})$. In every time period, all smart meters report their consumption to the aggregator V, who computes the sum of all consumption values $a_t = \sum_{i=1}^n e_{i,t}$ and finally reports a_t to ES. We remark that we do not model further knowledge of the ES explicitly, yet consider background knowledge of any malicious adversary implicitly through the metric proposed in the next subsections.

3.3.2 *Requirements of privacy notions for aggregation in the smart grid*

To assess the privacy protection offered by aggregation schemes in the smart grid, we identify the following requirements. A privacy

framework that allows to measure privacy leakage in aggregation schemes, should

- provide a strong formalism that allows reasoning about the provided privacy level, e.g., should allow to compute bounds; and should preferably
- allow to reason about practical attacks, i.e., it should be possible to show that these (with a certain probability) will fail.

Moreover, for a study of the trade-off between utility and privacy of the aggregated data, such a privacy framework should:

- provide an adequate adversarial modeling. Hence, it should consider a powerful adversary. Yet, the adversary's power should not be overestimated in order to achieve realistic assessments and to maximize utility.

Achieving an adequate modeling of the adversary, especially its background knowledge, which defines its strength, is a challenging task, which is discussed in more detail in the next subsections.

3.3.3 Smart grid privacy model

We define privacy for aggregation schemes using an indistinguishability notion. The core idea is to define privacy as the hardness to distinguish two load profiles known to the adversary in an aggregate. Informally speaking, the better the adversary in distinguishing profiles in aggregates, the weaker the privacy protection of individual households is in the aggregate. The strength of such a game based privacy notion is that it allows the modeling of arbitrary adversarial background knowledge, enabling us to model realistic and powerful attackers.

3.3.3.1 Formal Privacy Game

The basic game is illustrated in Figure 3.1. First, challenger and adversary agree on a load profile generator E_{gen} , the number of smart meters in the aggregate m , and the load profiles' length l . E_{gen} can either be a set of load profiles, e.g., from a real world consumption data set, or a sampling function that samples (realistic) load profiles from a probability distribution. After the initial setup phase, the adversary chooses (or samples, as described in the next paragraph) two load profiles \hat{e}_0 and \hat{e}_1 of length l from E_{gen} , which are then sent to the challenger. The challenger draws a random bit $r \in \{0, 1\}$, samples $m - 1$ further load profiles $\hat{e}_2, \hat{e}_3, \dots, \hat{e}_m$, and computes their aggregate $\hat{e}_a = \hat{e}_r + \hat{e}_2 + \dots + \hat{e}_m$. The aggregate is sent to the adversary who computes a decision function $f_{dec}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ that returns a bit

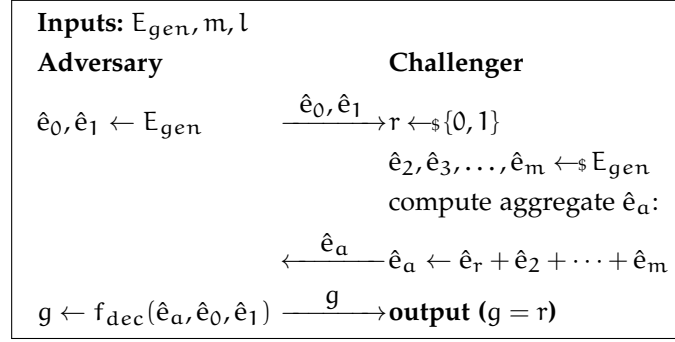


Figure 3.1: Basic privacy game for aggregation schemes (AggG) in the smart grid.

$g \in \{0, 1\}$, representing the guess whether \hat{e}_0 or \hat{e}_1 is contained in the aggregate. On a correct guess, the challenger outputs true, and false otherwise. We refer to the game as privacy aggregation game (AggG). The privacy of an aggregation scheme can be measured by the chances of an adversary in winning AggG. As in [42], we formally define the advantage of an adversary A for a given load profile generator E_{gen} , a number of smart meters m and load profile length l as the advantage over random guessing:

$$\begin{aligned} \text{Adv}_{\text{AggG}}^A(E_{gen}, m, l) = \\ |\Pr[\text{AggG}^A(E_{gen}, m, l, r = 0) = 0] \\ - \Pr[\text{AggG}^A(E_{gen}, m, l, r = 1) = 0]|. \end{aligned}$$

3.3.3.2 Practical Privacy Notion

Assuming an adversary with an optimal decision function, the outcome of one instance of the privacy game mainly depends on two aspects. Namely, it depends on the load profiles *chosen by the adversary* and the load profiles *sampled by the challenger*. For example, assuming two load profiles with very distinct (visual) shape chosen by the adversary and load profiles with a flat shape sampled by the challenger, these distinct shape of the chosen load profiles may also become visible in the aggregate and allows a decision with high certainty. Thus, the adversary's advantage in the privacy game noticeably depends on the load profile generator E_{gen} , namely, how distinct the generated load profiles are and how these are distributed, as the advantage is computed over all possible aggregates.

An adversary A maximizes its advantage by choosing load profiles that are the most distinct. Computing the maximum possible advantage allows to determine bounds on the privacy leakage and resembles the scenario for the worst case consumer with a very distinct energy consumption. We refer to this advantage as $(\text{Adv}_{\text{AggG}}^{A, \max})$. However, this notation might overestimate the privacy leakage for the average consumer, whose consumption is more similar to the average

energy consumption of other consumers. Therefore, we introduce a second interpretation of AggG, which is the average advantage over all combinations of load profiles that can be chosen by A ($\text{Adv}_{\text{AggG}}^{A, \text{avg}}$).

3.3.3.3 General Applicability

Revisiting the requirements of a privacy metric for aggregation, we observe that the AggG provides a strong formalism. Moreover, the applied indistinguishability notion is powerful, as it allows to model arbitrary, yet realistic background knowledge (load profiles are chosen from E_{gen}) of the adversary.

To illustrate the applicability of AggG, we consider the following exemplary privacy violation. The question whether it is possible to infer from a given aggregate that a consumer is at home during daytime can be modeled in AggG by choosing a load profile representing this consumption as \hat{e}_0 and a different typical load profile where the consumer is not at home as \hat{e}_1 . A significant advantage in AggG indicates that a malicious energy supplier is able to answer this question with some certainty. A further practical attack, which can be modeled with the AggG is illustrated in Section 3.5, where we show that individual appliances can be detected in an aggregate with their signature.

On a first glance the game based definition with the precise knowledge of \hat{e}_0 and \hat{e}_1 might seem as overestimating the adversaries capabilities. However, in practice energy suppliers have access to a significant amount of external information that can be very close to the knowledge of precise load profiles. For example, suppliers have knowledge about:

- households contained in an aggregate (technical requirement for most schemes),
- past load profiles of all aggregators,
- current and past monthly billing information for every smart meter and specific time charges,
- weather conditions, etc.

Moreover, we note that the adversary A in the AggG is given almost no background information on the energy consumption of the other households contained in the aggregate. A only knows that the aggregate is sampled from a subset of realistic load profiles. In practice, it is not unreasonable to assume, that a malicious supplier has further background information, e.g., to the average power consumption of multiple households contained in the aggregate, because these are also customers that periodically report their consumption for billing purposes. Moreover, as it has already been explored in the domain of location privacy, simple usage patterns and routines of individuals can be used to de-anonymize aggregates.

Due to these reasons, we consider the proposed metric as well suited to show which aggregation sizes are insufficient and risk the loss of privacy.

3.3.3.4 Further Application - Membership Disclosure

To further illustrate the versatility of the proposed metric, we describe how the indistinguishability based notion can be used to evaluate membership disclosure, i.e., to answer the question whether household x is contained in an aggregate or not. This can be evaluated by adapting the AggG, such that the adversary only samples one load profile $\hat{e}_0 \leftarrow E_{gen}$ and the challenger samples the other profile $\hat{e}_1 \leftarrow E_{gen}$, which is consequently unknown to the adversary. The rest of the game can be left unchanged, and the adversarial advantage is computed as the advantage over random guessing, whether \hat{e}_0 or \hat{e}_1 is contained in the aggregate.

We note that the membership game is at least as hard as the indistinguishability game. Given an adversary that can win the membership game, we can construct an adversary that is able to win the indistinguishability game with the same advantage. In order to decide which of two known profiles has been used in the aggregate of the indistinguishability game, an adversary could use the membership distinguisher to decide whether \hat{e}_0 is contained in the aggregate or not. The probability that \hat{e}_0 is in the aggregate is the same in both games. Hence, the adversary's advantage is identical to the advantage of the adversary in the membership game. However, the inverse reduction is impossible. Assuming that all but one load profiles, which can be generated from E_{gen} , are flat, i.e., constant, then an adversary in the indistinguishability game has roughly twice the advantage to observe the non-constant load profile than an adversary for the membership game, who only gets to see a single load profile. Thus, the membership game is strictly harder than the indistinguishability game. However, we observe that it is possible to construct a practical heuristic $f_{dec}^{mem}(\hat{e}_a, \hat{e}_0)$ for the membership game, given a decision function $f_{dec}^{ind}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ for the indistinguishability game. Even without access to \hat{e}_1 , an adversary in the membership game can repeatedly, i.e., a number of times k , invoke $f_{dec}^{ind}(\hat{e}_a, \hat{e}_0, \hat{e}_r)$ with a new randomly sampled \hat{e}_r . Using a majority voting $\sum_{i=0}^{k-1} f_{dec}^{ind}(\hat{e}_a, \hat{e}_0, \hat{e}_r) > k/2$, the adversary A decides whether \hat{e}_0 is in the aggregate. We give an experimental evaluation of this privacy question and heuristic in Section 3.5.

3.4 METHODOLOGY

In this section we first introduce the datasets under study. Then, we display our methodology followed in the experimental section and present the decision functions used to test the aggregates.

3.4.1 Smart grid datasets

To identify aggregation sizes that provide sufficient privacy, we apply the privacy game on multiple real world energy consumption datasets. In Table 3.1, an overview of the datasets used in this chapter is given. The datasets have mainly been made available for energy disaggregation research. To the best of our knowledge, these are the largest publicly available datasets regarding the number of load profiles. We observe that the datasets have different geographical origins, as well as different measurement set ups, e.g., resolutions. Moreover, we remark that the datasets use different types of power measurement including active, reactive and apparent power. Therefore, in most case studies we distinguish between datasets and study them separately.

Some datasets, e.g., Dataport and UMASS, contain several hundreds of households, whereas others, e.g., AMPds, focus on a single household for a large period of time. Unfortunately, only the Dataport and GOVAU dataset contain consumption data for more than 6 smart meters over multiple days.

Furthermore, most datasets require preprocessing, as they contain up to 10% incomplete or unusable (e.g., NAN) load profiles due to the experimental nature of energy consumption recording [95]. We consider a load profile to be complete if at least one sample is recorded in every sampling period required for a case study. Incomplete load profiles have been removed from all studies. The difference in the number of load profiles between complete and incomplete data is shown in Table 3.2. Note that the number of (complete) load profiles for each building in the same dataset may differ, therefore the total number of load profiles is given.

Table 3.1: Datasets used for the analysis. Presented are the geographical origin, the number of households measured in each dataset, the average number of load profiles that have been recorded for each household, and the sampling resolution.

Name	Origin	Households	LPs/Hh	Resolution
Dataport [90]	US	707	647	15 min
Redd [95]	US	5	7	1 s
AMPds [110]	Canada	1	726	1 min
ECO [40, 94]	Switzerland	6	192	1 min
UCI [105]	France	1	1358	1 min
GOVAU [81]	Australia	31	406	30 min
UMASS [38]	US	376	1	1 min

3.4.2 Evaluation approach

To identify an aggregation size that protects the consumer's privacy, an implementation on the privacy game was created. To handle most of the datasets, we rely on the NILMTK framework [39], which has been developed to study energy disaggregation algorithms (NILM). NILMTK provides converters for most of the aforementioned datasets into a consistent data representation. The adversary is modeled in the form of a decision function f_{dec} that decides between two chosen load profiles \hat{e}_0 and \hat{e}_1 . Different decision functions, which use a variety of heuristics, are introduced in the next section. For all case studies presented in Section 3.5, we apply the following algorithm:

1. For the analysis, we choose a dataset, an aggregation size m , a temporal resolution σ (the sampling frequency, e.g., $\sigma = 15$ min), an adversarial strategy (decision function f_{dec}), and a number of iterations N (e.g., $N = 5000$).
2. Then, the algorithm is reading the dataset. A dataset consists of multiple households with continuous load samples over one or multiple time periods. The load samples are grouped in load profiles of fixed start and end time. If not stated otherwise, each load profile starts at midnight with a duration of 24 hours in all experiments.
3. Next, the algorithm removes all incomplete load profiles, i.e., load profiles that do not have at least one load sample per sampling period.
4. If the input dataset is more granular than the chosen resolution, it reduces the resolution of all load profiles, by temporal aggregation of consecutive load samples.

Table 3.2: The number of buildings in each dataset that have at least one complete load profile and the total number of (complete) load profiles per dataset for a sampling resolution of 15min. The fraction of the usable against the total number of load profiles is displayed.

Name	Buildings		Load profiles		
	complete	total	complete	total	usable
Dataport	707	729	458048	474523	96.52%
Redd	6	6	53	236	22.45%
Ampds	1	1	726	730	99.45%
ECO	6	6	1196	1337	89.45%
UCI	1	1	1405	1440	97.56%
GOVAU	31	31	12606	12917	97.59%
UMASS	377	377	367	377	97.34%

5. The algorithm then selects two different households from the dataset uniformly at random. From the two households, it samples one load profile for each household. The sampled load profiles are labeled as \hat{e}_0 and \hat{e}_1 . This process ensures that even though different households might have a different number of load profiles, all households are represented equally in the result.
6. Analogously, it selects $m - 1$ load profiles from the remaining households. It samples a random bit $r \in \{0, 1\}$ and the $m - 1$ load profiles are summed up and added to \hat{e}_r to create an aggregated load profile \hat{e}_a .
7. The decision function f_{dec} is evaluated on \hat{e}_a, \hat{e}_0 and \hat{e}_1 .
8. If f_{dec} decided correctly (i.e., $f_{dec}(\hat{e}_a, \hat{e}_0, \hat{e}_1) = r$) a correct guess is recorded.
9. Steps 5-8 are repeated N times. Afterwards, the adversarial advantage values are computed as:

$$\text{Adv}_{\text{AggG}}^{f_{dec}, \text{avg}}(m) = \left| \frac{\text{correct guesses}}{N} - 0.5 \right| \cdot 2.$$

3.4.3 Decision functions

For two given load profiles \hat{e}_0 and \hat{e}_1 , an adversary in AggG has to decide which of the two is more likely contained in the aggregated load profile \hat{e}_a . In practice, finding an optimal decision is a hard computational problem, as an optimal distinguisher has to decide according to the maximum likelihood over all possible combinations of load profiles. Therefore, we focus on studying four heuristics and show in Section 3.5 that the described (comparably simple) heuristics are sufficient to identify load profiles in the aggregate. For better comparison, the aggregated load profiles are first normalized by the aggregation size: $\hat{e}_a \leftarrow \hat{e}_a / m$. The chosen decision functions are based on i) the *mean squared error (MSE)*, ii) the *Pearson correlation*, iii) *peak detection* and iv) a *combined method* based on Pearson correlation and peak detection. These heuristics have been chosen, as they all allow to measure a distance between two time series and follow different approaches.

In i) the MSE is computed as the pairwise squared difference between load samples, hence, $f_{dec}^{\text{MSE}}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ decides for \hat{e}_0 only if

$$\text{MSE}(\hat{e}_0, \hat{e}_a) < \text{MSE}(\hat{e}_1, \hat{e}_a).$$

The Pearson correlation also considers the trend of the compared load profile and ii) is decided by the higher correlation, hence, the function $f_{dec}^{corr}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ decides for \hat{e}_0 if

$$\text{corr}(\hat{e}_0, \hat{e}_a) > \text{corr}(\hat{e}_1, \hat{e}_a).$$

In iii) the relative peaks of each load profile $\hat{e}_a, \hat{e}_0, \hat{e}_1$, are determined and $f_{dec}^{peak}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ decides according the most common peaks between \hat{e}_0 and \hat{e}_a , or \hat{e}_1 and \hat{e}_a . Peak detection is a promising approach, as it considers the most significant features of a load profile that (in our expectation) could also be visible in an aggregate. For the peak extraction we follow a simple approach, where a window around every sample of length ± 1 (a resolution unit) is selected. If a sample has a value higher than its neighbors it is considered as a peak. For illustration, in Figure 3.2 a slice of an exemplary load profile \hat{e}_0 and an aggregated load profile \hat{e}_a , the identified peaks, and the windows of size three around each peak are shown. In the shown slice the load profiles only share one peak at 01:00 o'clock.

The decision function iv) combines peak detection and correlation with the idea that the shape of the load profile surrounding the peaks carries more information than the peak itself. Therefore, in iv) all peaks of \hat{e}_0, \hat{e}_1 and \hat{e}_a are computed. Then, the union of the peaks between of \hat{e}_0 (\hat{e}_1) and \hat{e}_a is formed. Afterwards, the Pearson correlation is computed for a surrounding window of a fixed length of samples around every peak, e.g. We identified a window of ± 5 (i.e., windows size is 11) as the best heuristic for 15 minute readings (a detailed analysis on the window size is given in the next section). The decision function $f_{dec}^{comb}(\hat{e}_a, \hat{e}_0, \hat{e}_1)$ decides according to the higher *mean correlation* between all windows of \hat{e}_0 and \hat{e}_a or \hat{e}_1 and \hat{e}_a .

3.5 CASE STUDIES

To analyze the privacy protection offered by aggregation schemes, we perform multiple case studies. First, we show for multiple datasets that the simple decision functions are sufficient to identify load profiles within aggregates of sizes ranging from two to hundreds of buildings. Moreover, we study the impact of temporal resolution, load profile length and daytimes on the distinguishing advantage. Then, we show that single appliances can be detected in aggregates consisting of load profiles from multiple households. Finally, we investigate membership disclosure in aggregates.

How effective are decision functions in identifying load profiles in an aggregate? We evaluate the effectiveness of the proposed decision functions by comparing them in the privacy game over $N = 5000$ simulations with different power measurements and time resolutions. A decision function is effective, if the advantage over random guessing

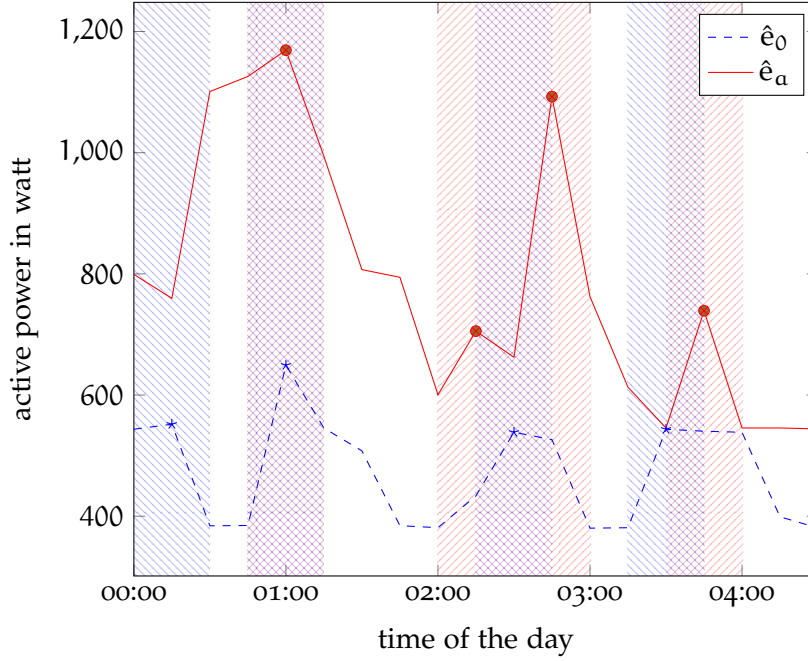


Figure 3.2: Shown is a slice of two exemplary load profiles (\hat{e}_0 , \hat{e}_a) and marked are the identified peaks of each load profile. Moreover, a window of size 3 is drawn around each load profile, which has been used to identify the peaks (cf., Section 3.4.3). The time resolution is 15 minutes.

is significant. The goal of the decision functions, as described in Section 3.4.2, is to identify the correct profile (\hat{e}_0 or \hat{e}_1) contained in the aggregate. First, we compare the average advantage of all proposed decision functions on the Dataport dataset with a sampling resolution of 15 minutes, shown in Figure 3.3. We observe that all heuristics can identify the correct load profile for small aggregation sizes with significant advantage. More precisely, for only 2 load profiles, all methods have an advantage of more than 75%. The Pearson correlation and peak detection heuristics perform similar over all evaluated aggregation sizes, whereas the proposed combination is the most powerful distinguisher. For aggregation sizes larger than 10, the advantage is more than twice higher, than the best advantage of the other three heuristics.

As already described in Section 3.4.3, the combined method computes the Pearson correlation for a window of load samples around all detected peaks. The window size, which influences the distinguishing advantage, is empirically evaluated in Figure 3.4. Plotted is the averaged advantage for different window sizes for the combined method on the Dataport dataset for three different sampling resolutions (15 min, 60 min, and 120 min) over aggregation sizes from 2 to 30. We observe that the best results are achieved for a moderately sized window, e.g., 10 load samples for a 15 minute reading. Moreover, we

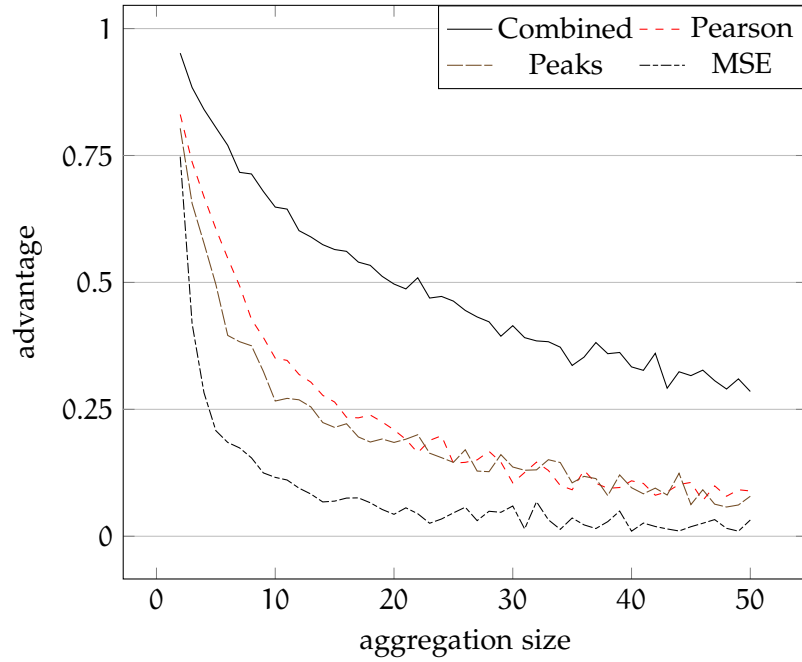


Figure 3.3: Comparison of the four decision functions, based on averaged adversarial advantage for different aggregation sizes (Dataport, 15 min resolution).

observe that a more granular sampling resolution requires more load samples to be contained in the window to achieve the best advantage.

Which parameters influence the privacy game?

Dataset dependency. The results of an empirical analysis commonly depend on the dataset used. To show that the difference in adversarial advantage is rather small between the datasets, we compare the distinguishing advantage between the datasets Dataport and the other two largest datasets (GOVAU and UMASS), for the combined decision function in Figure 3.5. We observe that the power consumption in UMASS is noticeable more distinguishable by the combined decision function than the GOVAU and Dataport dataset, which share a very similar (in-)distinguishability for increasing aggregation sizes.

Furthermore, we can illustrate a similar behavior of all four decision functions on a union of all load profiles from all datasets. To sample a load profile in this experiment, we first sample a dataset, then a household (uniformly among the dataset) and then a load profile (uniformly among the household). This guarantees an equal representation of datasets and households. We acknowledge that consequently some load profiles have more impact on the results than others, unfortunately the limited number of large datasets does not allow for a better experimental setup. The distinguishing advantages in this experiment are shown for different groups of aggregation sizes in Figure 3.6. We observe a similar distinguishing advantage as when studying datasets

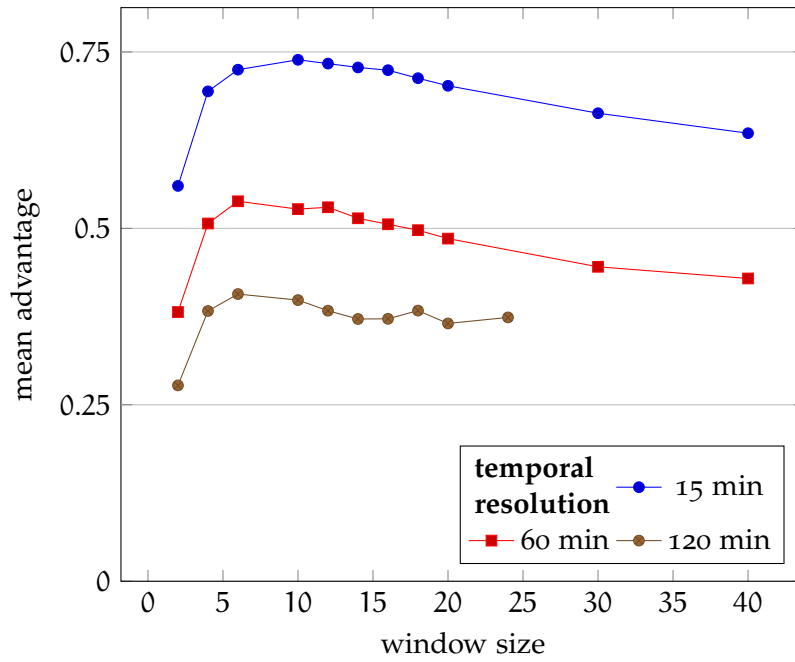


Figure 3.4: Evaluating the window parameter of the Combined decision function (Dataport, $m = \{2, \dots, 30\}$).

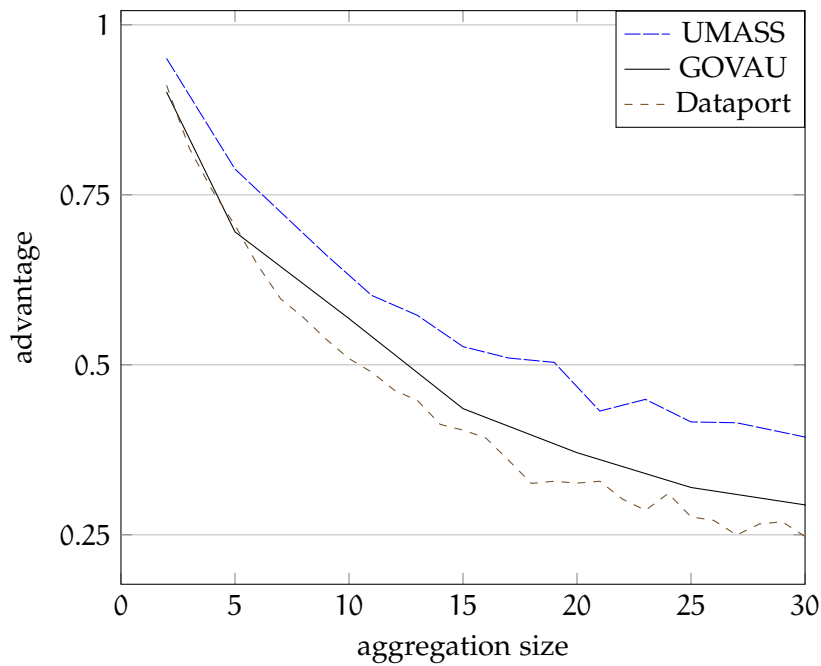


Figure 3.5: Comparison of the averaged distinguishing advantage between the UMMASS, GOVAU, and Dataport dataset for different aggregation sizes, when using the combined decision function (30 min resolution).

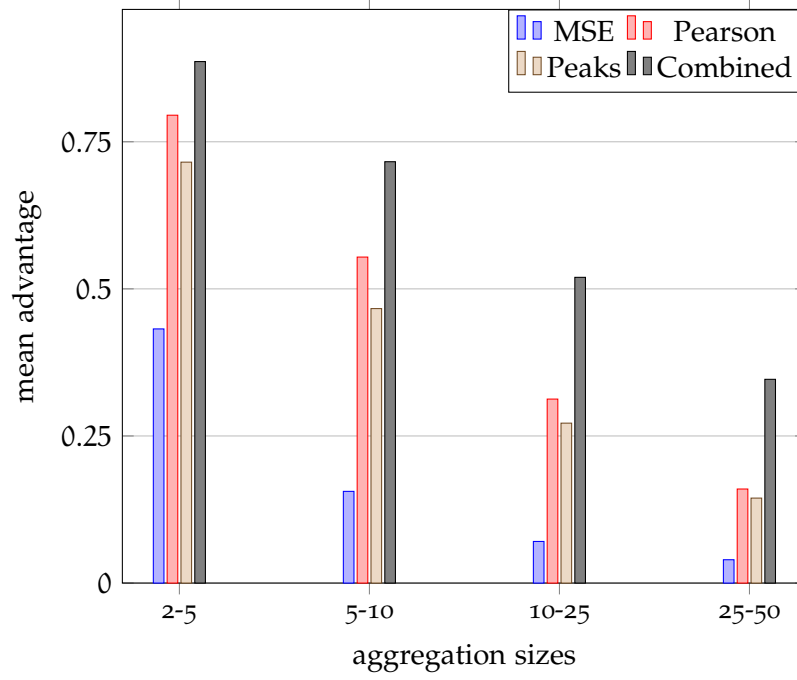


Figure 3.6: Comparison of the decision functions based on the averaged adversarial advantage, using different aggregation sizes, using profiles from all datasets (30 min resolution).

independently. Moreover, as before the combined decision function outperforms the others in every scenario, and hence will be used as the main decision function in the remainder of this section.

In summary, the datasets show diversity in their load profiles, which is also visible in the AggG. However, the difference in the results between the datasets (also influenced by the empirical nature of our approach) is only marginal when deriving qualitative statements on the individuals' privacy.

Temporal resolution. The rate at which each smart meter reports its values, is a crucial factor for privacy. More frequent reports enable NILM algorithms to work with higher accuracy and extract more information. Hence, we expect that higher sampling resolutions are less privacy friendly. In Figure 3.7, we present the advantage for different aggregation sizes with different sampling resolutions when using the combined decision function on the Dataport dataset. Clearly, the advantage is higher for more frequent reports and smaller aggregation sizes. In aggregations with 10 or more load profiles ($m \geq 10$), the advantages differ only by a small factor independent of the temporal resolution. When only two households are aggregated, a significant advantage of 50% is observed for a temporal resolution as low as 4 hours. This confirms what we intuitively expected, namely that distinguishability increases with more frequent reports. In addition, it is

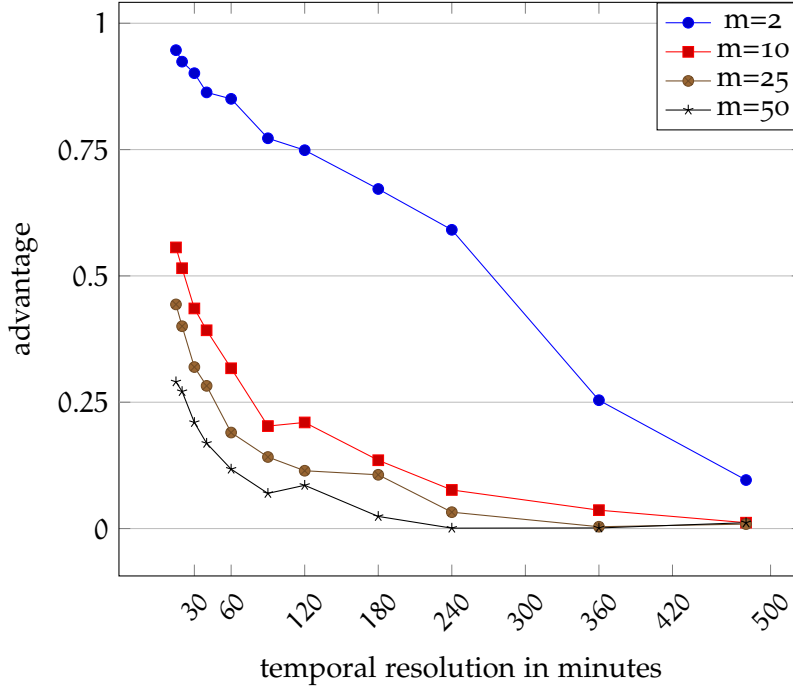


Figure 3.7: Comparison of the impact of different temporal resolutions when using different aggregation sizes m . Measured is averaged advantage using the combined heuristic (Dataport).

clear from measurements that the aggregate of two load profiles is not enough to provide privacy, even for a very low sampling resolution.

Influence of different daytimes. In previous evaluations, we studied load profiles of 24h length. In this section, we examine if *different daytimes* affect the model's accuracy. The load profiles of the Dataport dataset were split in four parts according their daytime. Those were, night time (0:00-6:00), mornings (6:00-12:00), afternoons (12:00-18:00) and evenings (18:00-24:00). We study aggregation sizes ranging from 2 to 50 households, and fix the sampling resolution to 15 minutes. We performed 5000 simulations of the privacy game for each period. In Table 3.3, the average advantage for the four different daytimes, as well as for the whole day is presented. In summary, we observe only marginal differences between the different daytimes, but as expected, a 24 hour load profile allows for better distinguishability than isolated daytimes.

How many households are required to achieve privacy? Bigger aggregation sizes lead to better privacy for individual households. However, an arbitrary increase in aggregation size defeats the purpose of smart meters, which should be able to monitor and predict the consumption in order to distribute energy more efficiently. Thus, an upper bound exists on how many households should be in an aggregate report in order for the smart grid to retain some utility. Unfortunately, we have no (reasonable) measure of utility, yet we can identify a marginal util-

Table 3.3: Average advantage when distinguishing load profiles of 6h length, using the Dataport dataset (15 min resolution), compared with the advantage when distinguishing a load profile of 24h length.

Daytime	$m = 2$	$m = 5$	$m = 10$	$m = 30$	$m = 50$
Night	0.811	0.574	0.411	0.233	0.184
Morning	0.836	0.603	0.436	0.232	0.157
Afternoon	0.792	0.566	0.418	0.237	0.157
Evening	0.800	0.566	0.410	0.234	0.152
Day - 24h	0.947	0.793	0.634	0.396	0.29

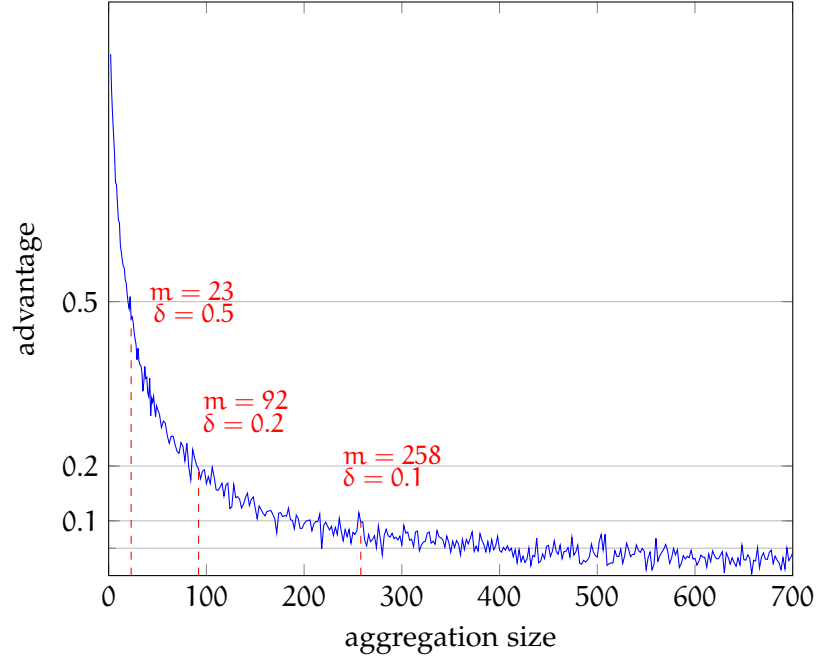


Figure 3.8: Adversarial advantages achieved with the combined decision function for different aggregation sizes m . Marked with δ are interesting advantage levels (Dataport, 15 min resolution).

ity on the privacy protection. Applying the AggG with the combined heuristic on the Dataport dataset, which provides the largest number of load profiles and households, for aggregation sizes of up to 700, we can infer, which aggregation size is needed to achieve a certain level of privacy (distinguishing advantage over random guessing) shown in Figure 3.8 for a 15-minute sampling resolution. We denote with m the size of the aggregate and with δ the average adversarial advantage. The shape of the curve can be used to analyze the marginal utility. The curve is very steep up to a privacy level of $\delta = 0.5$, which is reached in the experiment with an aggregation size of $m = 23$. At a privacy level of $\delta = 0.2$ ($m = 92$) the curve significantly starts to flatten out with only marginal improvements in privacy after $m = 200$.

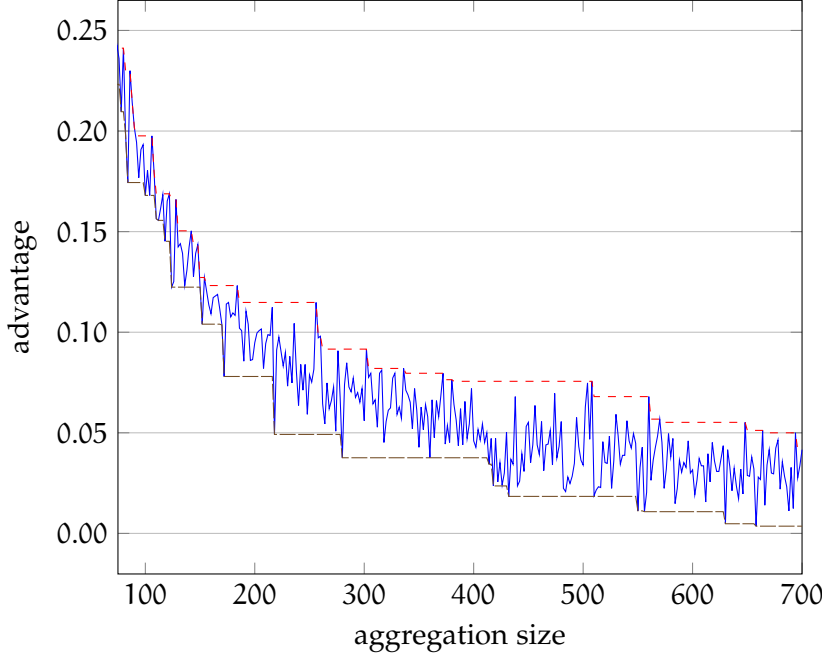


Figure 3.9: Illustration of the imprecision in the experimental computation of the distinguishing advantage for aggregation sizes $m > 75$ over simulations with $N = 5000$ runs (Dataport, 15 min resolution).

We remark, that results for the distinguishing advantage for larger aggregation sizes, e.g., above 100, should be studied with a grain of salt. Even though, each data point is computed via a simulation over 5000 trials, it contains a noticeable error for larger aggregation sizes, which is illustrated in Figure 3.9. We observe that similar aggregation sizes can show a noticeable variance in the distinguishing advantage.

Are particular appliances detectable in an aggregation scheme? Specific appliances create unique patterns in load profiles. NILM algorithms can extract specific devices' usage in single households, by detecting those patterns. In order to examine if specific devices can be detected in the aggregation, we adjust the adversary's choices in the AggG. The first load profile \hat{e}_0 is sampled from the dataset, the second load profile is generated by subtracting from the individually measured load profile \hat{e}_0 a single appliance. Hence, \hat{e}_0 and \hat{e}_1 only differ in the energy consumption of a single appliance.

Using the Dataport dataset, we study the AggG for different aggregation sizes and household appliances. In Figure 3.10, we present the average adversarial advantage over random guessing, for an electric furnace, a dish washer, a fridge and a stove. The results demonstrate that specific appliances, e.g., electric furnace, are detectable with significant advantage even in aggregates of size $m > 10$. As expected, the detection is more powerful when aggregation is small. Figure 3.11 presents the adversarial advantage when detecting various devices in the Dataport dataset, for aggregation sizes of $m = 5, 10$ and 25.

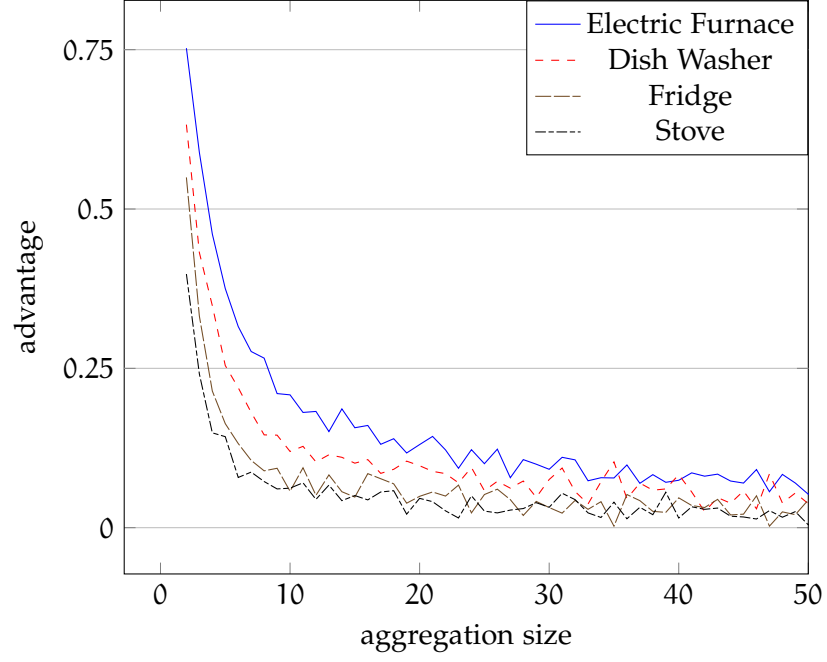


Figure 3.10: The detectability of single appliances in an aggregate. Shown is the distinguishing advantage for the combined decision function (Dataport, 15 min resolution).

Table 3.4: Characteristic properties of particular appliances (average values).

Appliance	Mean load(W)	Max load (W)	#Peaks	Daily uptime	Load/Peak
Dish washer	43.5	887.1	6.1	9%	504.5
Electric furnace	135.7	603.7	18.7	87%	305.8
Fridge	77.1	344.5	23.2	50%	143.9
Stove	55.0	1110.8	9.6	27%	440.7

To identify specific patterns that make a load profile (of an appliance) distinguishable in an aggregate, we study various properties of appliances in the dataset, namely: mean load (when switched on), maximum load, the number of peaks, as well as the daily uptime and average load per peak. The results are illustrated in Table 3.4. The correlation between the characteristic properties and the detectability, using the Pearson correlation between the properties and the advantage per aggregation size is depicted in Figure 3.12.

We observe that the detectability of an appliance shows the largest correlation with the maximum load, followed by the average load per peak. The correlation between average mean load and the detectability is significantly lower, while the properties average daily uptime and number of peaks are negatively correlated to the privacy level. To conclude, not only households but also individual appliances that

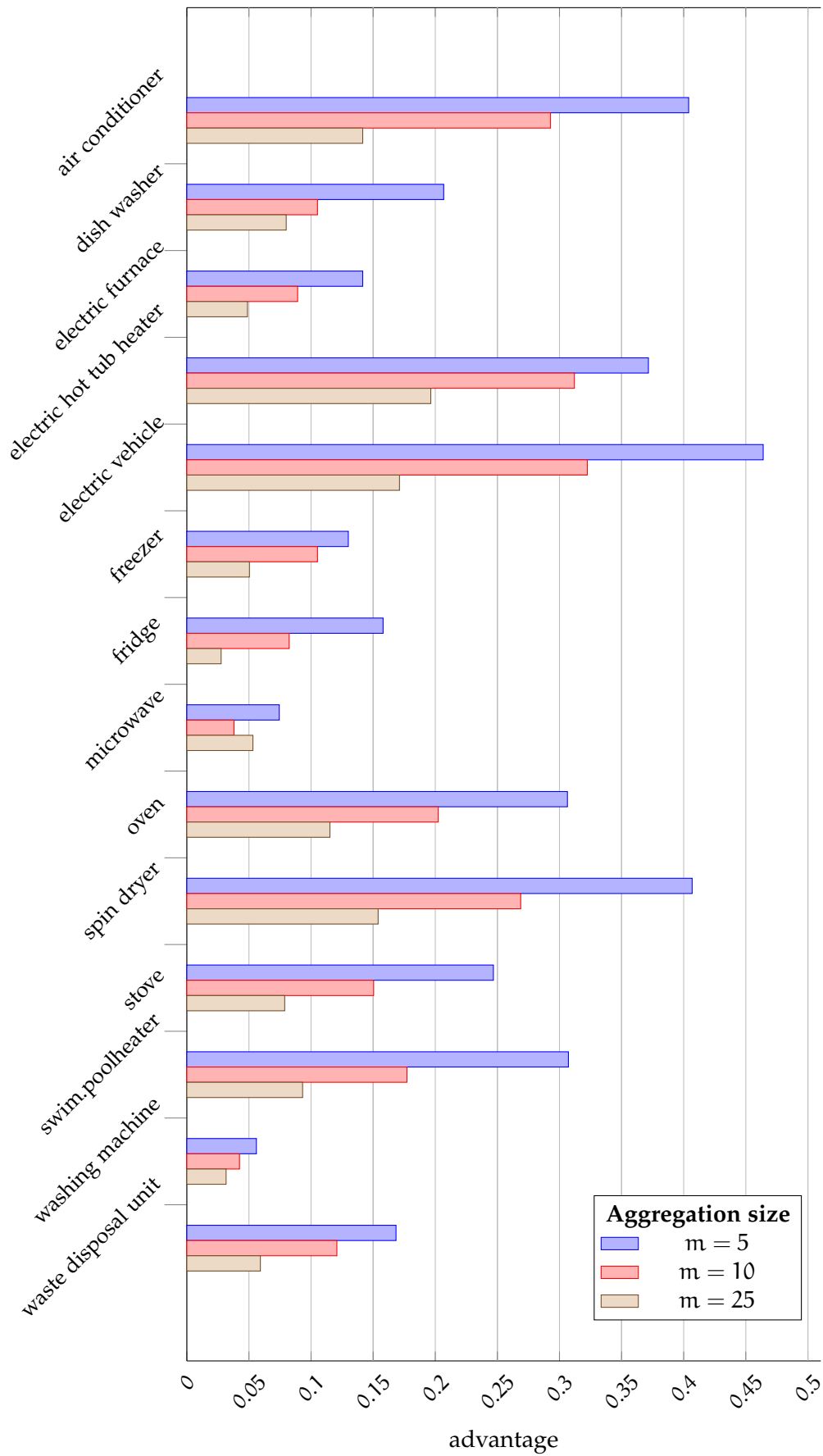


Figure 3.11: Average adversarial advantage for particular appliances. Three different aggregation sizes tested ($m=5,10,25$) (Dataport, 15 min resolution)

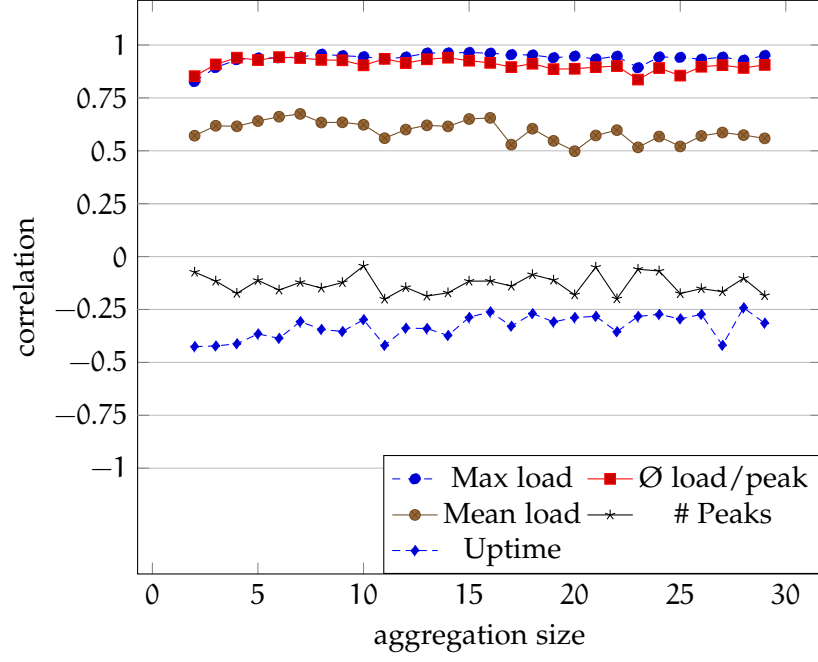


Figure 3.12: Correlation between characteristic properties and detectability of appliances for different aggregation sizes.

show consumption patterns with high peaks can be detected with minimal effort in aggregates of smaller size.

Can membership in an aggregate be disclosed using the same decision functions? A further application of AggG is outlined in Section 3.3, namely whether an adversary could identify the existence of a single profile in an aggregate rather than distinguishing two known profiles. Hence, shifting the focus from an indistinguishability notion to a membership disclosure question. As described, the privacy game has to be adapted in the following way; instead of having the adversary select two profiles and then try to distinguish which one is in the aggregate, he only samples one (\hat{e}_0). Then the challenger randomly samples a second one (\hat{e}_1), unknown to the adversary, and by flipping a coin decides which one of the two will be used in the aggregate sent to the adversary. The adversary has then to guess, whether this profile is part of the aggregate or not. Using a similar experimental setup as before, we studied this question for the Dataport dataset with aggregation sizes from $m = 2$ to 20 and $N = 1000$ simulations per aggregation size. Moreover, we used the decision functions as described in Section 3.3 with $k = 100$ iterations. In Figure 3.13 the advantage for correct answering the membership question with the help of the two most effective heuristics, i.e., peak detection and the combined method is presented. In addition, the advantage in the indistinguishability notion for the same aggregation sizes is given.

Even though the adversary has less power in this game, and consequently, the advantage decreases compared to the indistinguishability

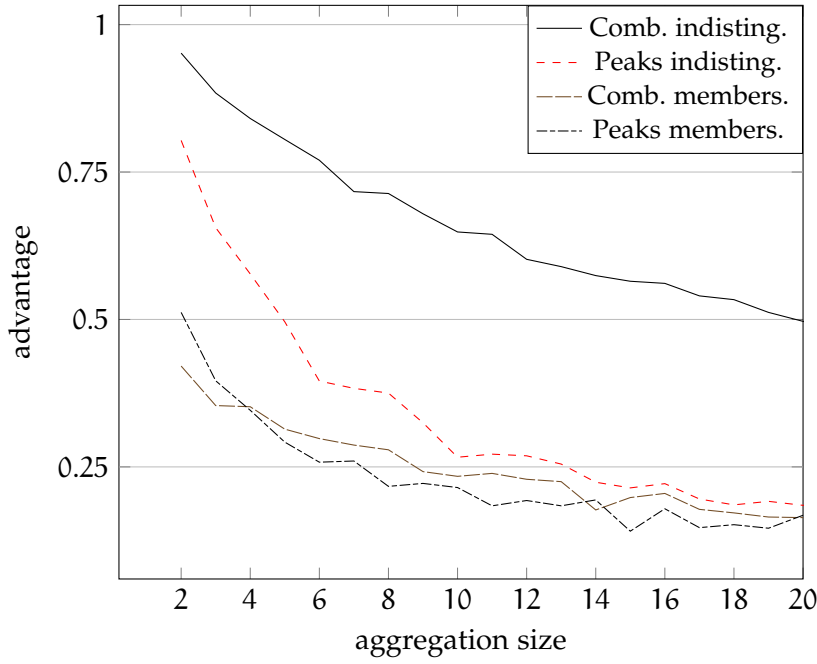


Figure 3.13: Comparison of the modified privacy game where the adversary has knowledge of only one load profile (membership disclosure) vs when he distinguishes two profiles (indistinguishability notion). For each experiment, the distinguishing advantage for the two most effective heuristics (peak detection and combined method) for aggregation sizes of up to 20 households is displayed. (Dataport, 15 min resolution).

game, we observe that the advantage remains significant for all aggregation sizes. A more surprising result is that that peak detection and the combined methods perform similarly for membership disclosure, in contrast with the cases previously examined. This again indicates that the peaks are the most robust feature to distinguish load profiles.

In summary, the AggG is very suited to also examine privacy under a different view point, i.e., membership disclosure, with small modification. Furthermore, even in the membership based privacy notion, very simple heuristics are able to achieve a significant advantage over random guessing for larger aggregation sizes.

3.6 DATASET ANALYSIS

Studying the detectability of individual load profiles with the help of the aggregation game, a question arises, whether a common universal load profile exists. The existence of a universal load profile could be used to only consider the relative changes to the universal load profile as privacy relevant and thus, demand a reformulation of the privacy game. Therefore, in this section, we first study the differences between individual load profiles and their average from the dataset. Second, to overcome the very limited availability of real world energy consump-

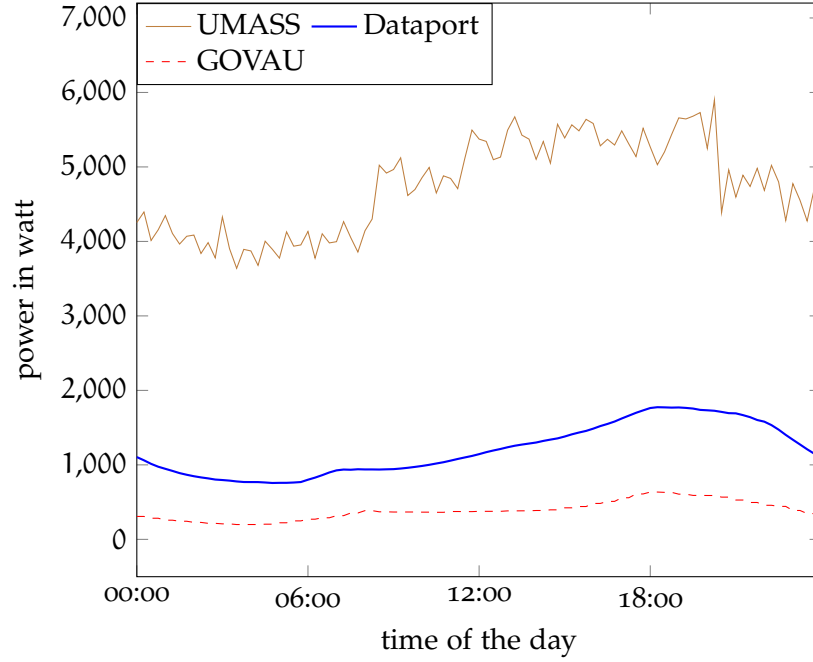


Figure 3.14: Mean load profiles of the datasets UMASS, Dataport, and GOVAU.

tion datasets, we study the applicability of load profile generators in privacy research for the smart grid.

Universal load profile. While datasets from different countries presumably differ in their average load profile due to differences in cultural and climatic preconditions, this does not apply to load profiles from similar climate zones and cultural environments. Unfortunately, the available data is insufficient to present an exhaustive analysis. Yet, when comparing the average load profile of the Dataport (707 households) and GOVAU (31 households) and UMASS (376 households), shown in Figure 3.14, similarities in shape can be identified. For example, comparatively low consumption values during night, and consumption peaks in the morning and evening hours are visible. We note that the different energy measurements (re-/active power) lead to noticeable differences in the individual consumption values and should therefore not be compared by their absolute value. Moreover, we observe significant more variance in the UMASS datasets, which only provides one load profile for every household.

Generally, we observe that individual load profiles can be quite distinct from the average of a dataset. This is illustrated in Figure 3.15, where the distribution of the mean squared error (MSE) between all individual load profiles and their average is illustrated for the Dataport dataset. Similarly, in Figure 3.16 the Pearson correlation coefficient between individual load profiles and the dataset's average is plotted for the GOVAU and ECO dataset. Even though, there is

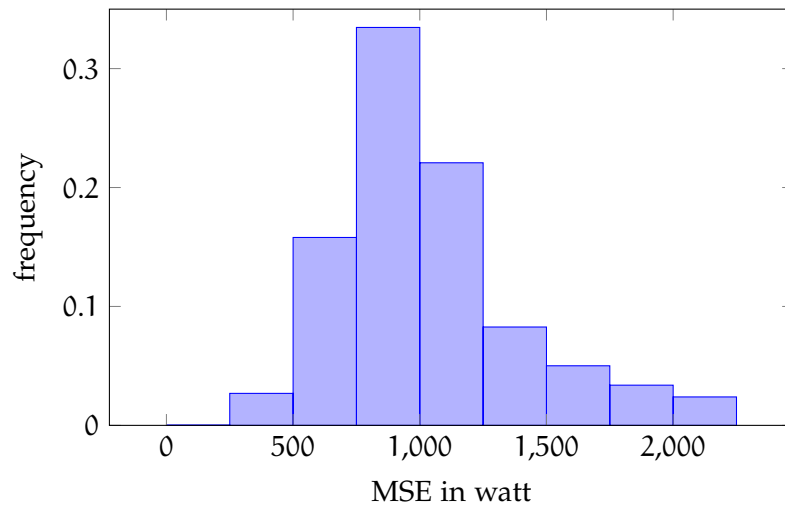


Figure 3.15: Distribution of Mean Squared Error between individual load profiles and the average load profile in Dataport.

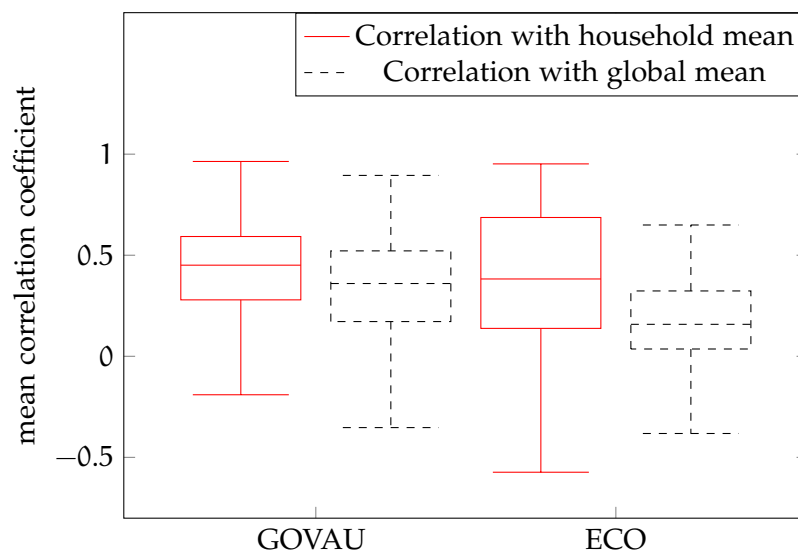


Figure 3.16: Distribution of correlation between single load profiles and the datasets (GOVAU, Eco) or households average.

a noticeable correlation to the datasets' averages, we also observe numerous outliers in both datasets, that are very different to the average. Thus, load profiles carry significantly more information than only small relative differences to the datasets average load profile. Moreover, Figure 3.16 also shows the correlation of each load profile to the household's average. This correlation is (expectable) higher than the correlation to the mean of the according dataset, yet also shows a significant variance between the load profiles from the same household.

In summary, we observe that (background) knowledge on the average load profile of a region or household could be used to improve the detectability of load profiles in an aggregate. Yet, the significant variance between load profiles illustrate that the protection of only small changes to an average is insufficient.

Load profile generators for privacy research. Load profile generators are tools that are able to simulate the energy consumption for a certain period based on an underlying model. The area of application ranges from studies concerning the effects of new technology on the energy consumption of households to forecasts like the determination of the national energy demand [143]. A natural question is if these generators are adequate for a privacy analysis and could be considered in subsequent work. In this work, we focus on studying energy consumption of individual households, and hence a load profile generator based on the so called bottom-up approach is promising. Bottom-up load profile generators aim at simulating the behavior of inhabitants of a household, that is modeled in the use of household appliances, e.g., cooking, heating or television or other activities. According to the simulated usage of appliances and (pre-recorded) appliance specific demand profiles, load profiles for households are generated. The model can be enhanced by external influences like temperature, holidays and geographic circumstances. Various bottom-up load profile generator have been proposed in [126]. Using the implementation of the *Loadprofile Generator* [28], we created a dataset containing 266 households with 365 load profiles per household, which is studied in the following paragraphs.

In Figure 3.17 the mean load profile of the generated dataset is shown. When comparing the dataset's mean to the mean load profiles of the datasets depicted in Figure 3.14 qualitatively, it is clear that all datasets share similarities and prominent features like the peaks at about 06:00 and after 18:00 can be found in both.

Similar to the previous analysis, we applied the privacy game to the generated dataset using a resolution of 15 minutes on aggregates of size $m = 2$ to 50. The results for all decision functions are illustrated in Figure 3.18. We observe that all decision functions show a similar behavior as on the real data sets, i.e., the combined decision function outperforms the rest, whereas the MSE shows the least distinguish-

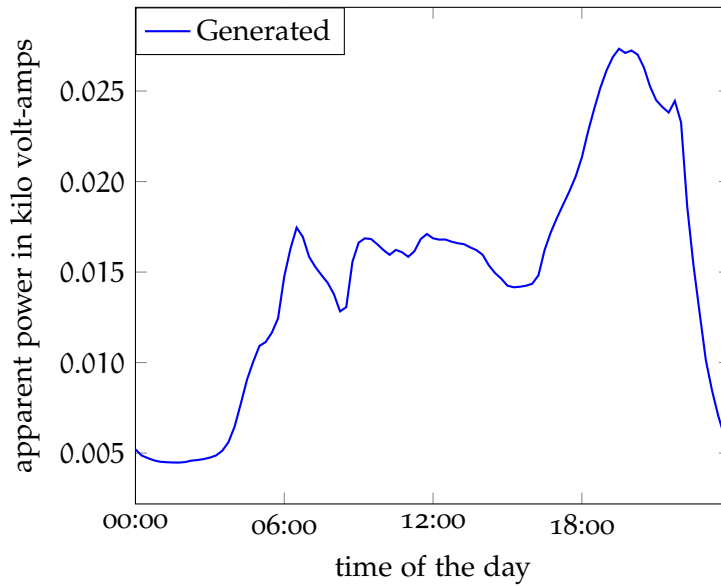


Figure 3.17: Mean load profile for the generated dataset consisting of 266 households.

ing advantage. For better comparison, we plotted the results of the combined decision function when applied on the Dataport dataset. Also here, we observe no significant differences (in the distinguishing advantage) between the generated dataset and the real dataset. Hence, we conclude for future privacy studies on real world data, generated datasets seem to be a very promising alternative to real world datasets, whose availability is very limited.

3.7 CHAPTER SUMMARY

In this chapter, we evaluated the privacy guarantees of aggregation in smart metering. Using realistic datasets and modeling privacy as an indistinguishability game, we showed that contrary to industrial insights, small and medium sized aggregates can leak information about users' activities. Furthermore, we evaluated various parameters of the aggregate that affect the privacy offered. We illustrated that less frequent reports improve privacy guarantees but do not completely eliminate the privacy danger. We further concluded that, we cannot argue about a universal profile of household electrical consumption, based on the datasets we studied. Last, we studied electricity consumption profile generators, and we showed that in the lack of real datasets, they can be a viable alternative.

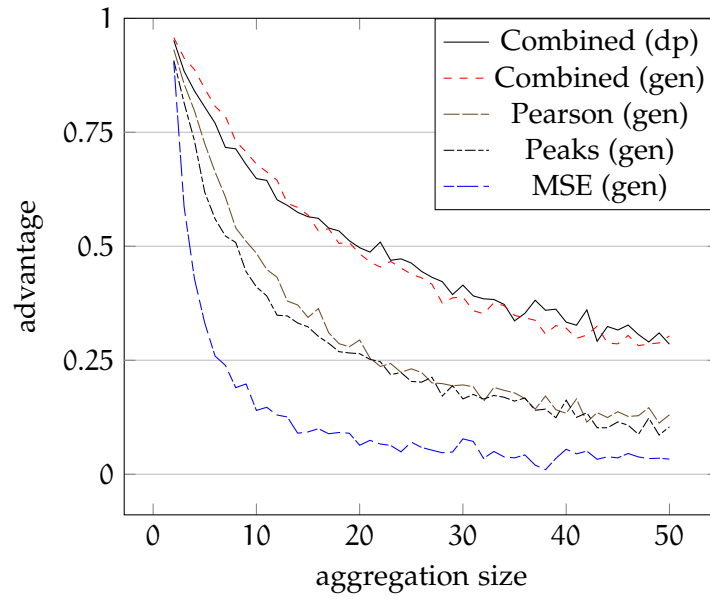


Figure 3.18: Comparison of the decision functions for the generated dataset (gen). For comparison, the advantage of the combined decision function applied to the Dataport (dp) dataset is also plotted (15 min resolution).

MEASURING PRIVACY IN HIGH DIMENSIONAL MICRODATA COLLECTIONS

Published Content

S. Boukoros and S. Katzenbeisser.

“Measuring privacy in high dimensional microdata collections.”

In: International Conference on Availability, Reliability and Security, 2017, pp. 15:1–15:8.

4.1 MICRODATA RELEASES AND PRIVACY

In this chapter, we focus on privacy on microdata collections. Microdata refers to collected statistical information about individuals; for example, the movies that someone has watched or the pages she likes on Facebook. This kind of statistical data regarding users is a lucrative asset for a variety of companies and organizations, because a plethora of information can be extracted, such as shopping trends, health or financial indicators, etc.

Microdata pose a privacy risk in case parts of databases are compromised, or even when datasets are published sanitized (i.e., anonymized) as inadequate procedures are typically used in the anonymization process, such as the simple removal of personal identifiable information, i.e., names, tax IDs, etc. In a demonstration to show that these procedures are not safe, a US politician was reidentified through published anonymized health records [145], while in [115] the authors demonstrated that half a million users of a popular movie database could be reidentified with high accuracy.

Whenever such personal data/preferences become publicly available, individuals are affected in a variety of ways. It is known that political orientation, sexual preferences, age, intelligence and other personal traits can be statistically inferred with high accuracy just by accessing someones online profile [96]. In addition, unexpected connections in social graphs, so called weak ties, can uncover identities and reveal sensitive data [132]. For example, if the online profiles of two users share some common interests, we can predict traits about one profile by having access only to the other. Additionally, a privacy breach has implications for future privacy [115]. In a hypothetical scenario where Alice’s identity has been revealed, she can no longer hide behind another online account with a different nickname. Even if

she creates a brand new account, her microdata will *always* identify her.

In this chapter we develop a novel tool for individuals' privacy assessment based on their microdata. Similar to the previous chapter, we measure privacy in terms of indistinguishability. However, there is a difference. In this chapter we approximate users' indistinguishability in terms of their uniqueness in a database. In a nutshell, how likely it is, that their preferences are unique and if someone knows them (or a subset of them that is unique), they can identify them in an pseudonymous (or anonymized) dataset. The tool approximates a user's privacy level and categorizes her in a privacy group ranging from a relatively safe privacy group, where many users have similar preferences, to a dangerous group where her data are uncommon and hence, not many people share them.

The proposed tool measures the risk of being reidentified based on a user's microdata, *prior to* sharing their data, based on two factors: the data the user wants to *disclose to the provider* and *published statistics about those data*. We consider the following scenario: Alice wants to use a service for music recommendation. However, Alice heard on the news about recent reidentification attacks on similar services and is worried about her online privacy. By using the proposed tool, she can obtain an estimation and a visual representation of the risks being reidentified based on her microdata (privacy level), without disclosing anything to the provider. With this extra information, she no longer has to trust the provider but can *consciously decide* whether she wants to contribute her data.

The tool is composed of two instances, one for the service provider and one for clients, and a privacy metric. On the provider's side, the tool clusters clients based on their microdata into privacy groups and publishes information regarding each group. In addition, statistics regarding the items' popularities are published. Such kind of statistics already exist in many services, usually in the form of "another x users liked this item". The users' instance measures the privacy level of a user, based on such published information by the provider. As a last step, the tool provides a visual and easily comprehensible result to the user.

Chapter outline

We proceed by reviewing related work in Section 4.2 and we describe the tool and the privacy metric in Section 4.3. In Section 4.4, we introduce the datasets used for evaluation, and present the experiments demonstrating that our tool can capture the reidentification risk of someone adequately, even without full access to the database. In Section 4.5 we describe the proposed visual way in which the privacy levels are presented to users.

4.2 RELATED WORK

4.2.1 *Reidentification attacks*

Various works have shown that combinations of attributes can be used for de-anonymization. Sweeney [144] conducted experiments on the 1990 US census summary data to determine if and how combinations of identifiers could uniquely identify people. It was shown that 87% of the US population can be identified based on their 5-digit ZIP code, their gender and their birth date. De Montjoye et al. [56] studied the uniqueness of individuals in an anonymized credit card transaction dataset using information from 1.1 million people collected over 3 months. Their experiments show that over 90% of the users can be uniquely identified by only 4 places where they pay with their credit cards. Furthermore, the deanonymization process becomes easier if the adversary knows the approximate price of a transaction. The odds of reidentification are also higher for women and for people with high income. Solomon et al. [141] investigated individuals' uniqueness in high dimensional social sciences datasets, collected for medical purposes. Their results indicate that only 3 attributes are enough to deanonymize every participant in the dataset. The rest of the attributes however, are rather similar indicating that aggregation techniques based on those 3 crucial attributes could mitigate the privacy risk.

Various reidentification attacks, mostly on publicly available sanitized datasets, demonstrate the risks involved in data publishing. Sweeney [145] was able to identify the profile of a US governor in an anonymized health database by combining it with a purchased voters registration list. His profile was the only one matching the fields of ZIP code, gender and birth date. Frankowski et al. [73] studied the problem of linking users' profiles between a private and a public database. More specifically, using a popular public movies discussion forum, they matched users to a private movie rating database using several different algorithms. Their results portray that relationships to items in sparse datasets can be used as quasi identifiers. Furthermore, they studied the effects of suppression and generalization mechanisms on the datasets. Their experiments suggest that the best strategy is addition of noise; mentioning movies that one has not actually rated in order to confuse the matching algorithms.

Narayanan and Shmatikov [115] presented new statistical de-anonymization attacks against microdata datasets. Their attacks rely on the similarities between a user's profile and the auxiliary adversarial information. In their work they showed how adversaries with little background information and not necessarily fully correct, can deanonymized microdata collections. By performing their attacks on an actual dataset published by Netflix, they illustrate that as few as 5-10 attributes per user can lead to a privacy breach for thousands of

users. Merener [113] elaborates on the algorithms used in [115]. His research focused on the generalization of the results under general and realistic assumptions, and presented an improved version of one of the algorithms. In addition, he highlighted the importance of a dataset's sparsity in the deanonymization success and the importance of rare items in users' distinguishability. He evaluates the theoretical findings by experimenting with an anonymized health dataset.

Al-Azizy et al. [32] survey data deanonymization techniques and cluster them according to they type of auxiliary information, and the structure of the datasets.

4.2.2 *Privacy quantification*

Various metrics based on data similarity have been proposed to measure privacy, independent of adversarial models. K-anonymity [145] is a property of statistical databases. It requires that groups of at least k entries have the same values for the quasi identifiers while all direct identifiers have been removed. Thus, an adversary cannot uniquely isolate an entry but has to choose from further $k-1$. In order for a database to satisfy k -anonymity, generalization or suppression mechanisms can be used. L-diversity [109] is an extension of k -anonymity. After classes with at least k entries have been formed, this definition aims to limit the diversity of all created groups. It is required that each class should contain at least l well represented sensitive values; for instance at least l distinct values should be published per class, or the most frequent ones. Another extension of k -anonymity is t -closeness [104]. It requires that the distribution of sensitive values in all classes should follow that of the general database otherwise, adversaries with auxiliary information on the global distribution of the sensitive values, could infer selected attributes. Hence, the distance of the distributions of each pair of classes should be smaller by a threshold. However, due to multidimensionality and sparsity in microdata publication, the effect of the latter anonymization approaches and privacy metrics might be inadequate [31].

Parra-Arnau et al. [124] proposed the use of the Kullback-Leibler divergence, an information theoretic quantity, as a privacy metric for profiles in personalized information systems. The motivation behind such a proposal is that the average user profile is the one that maximizes entropy. Hence, any deviation from such a profile can be viewed as privacy loss. Thus, users whose profiles deviate from the average profile are more probable to be targeted or classified by adversaries.

Differential privacy is an indistinguishability metric first proposed for the domain of databases [63]. When a dataset is differentially private, any result is guaranteed to be equally likely regardless of the weight of any individual entry in the database. In other words, the existence or not of an entry in the dataset, cannot affect the result

of a query more than a predetermined value. Differential privacy is usually achieved with the addition of noise in the data and comes in two settings; an interactive and an offline. In the interactive one, users make queries to the sanitized database until a privacy budget is fully spend. In the offline setting the data are sanitized once and published. A detailed analysis of the most used privacy metrics was done by Wagner and Eckhoff [152], while a survey of privacy preserving data publishing techniques was presented by Chen et al. [46].

Our contribution

This chapter proposes a tool for estimating users' privacy levels based on their shared information with their provider, and the popularity of their items. While we do not propose any new attacks, we use the notion of auxiliary information to simulate an attacker that tries to match this information against entries in the database. Contrary to most proposed privacy metrics, the one developed in this chapter does not require full access to the providers' databases and other people's profiles. However, due to that, the privacy result is an approximation of the users' actual privacy rather than an exact estimation.

4.3 PRIVACY ASSESSMENT MECHANISM

4.3.1 *Overview of the system*

An overview of the proposed system is presented in Figure 4.1. The purpose of the tool is to allow users to estimate their privacy level (i.e., their reidentification risk) before using an online service where microdata might be collected. In order to do so, the provider needs to create privacy groups, based on his existing clients, using our proposed metric (Figure 4.1: A). These groups consist of users whose risk of reidentification is on roughly the same level. In addition, the provider publishes statistics regarding the popularity of all items (Figure 4.1: B). On the client's side, the tool estimates the privacy score using those public statistics and his microdata (Figure 4.1: C and D). Finally, the user gets a visual representation of his privacy level, illustrating the privacy group in which he was categorized.

4.3.2 *System's internal logic*

4.3.2.1 *Privacy metric*

The tool requires a metric that can estimate individuals' privacy level locally, using minimal information from the provider. Hence, the need for a new privacy metric arises, as existing metrics used in databases are not applicable due to two reasons. First, we work with providers

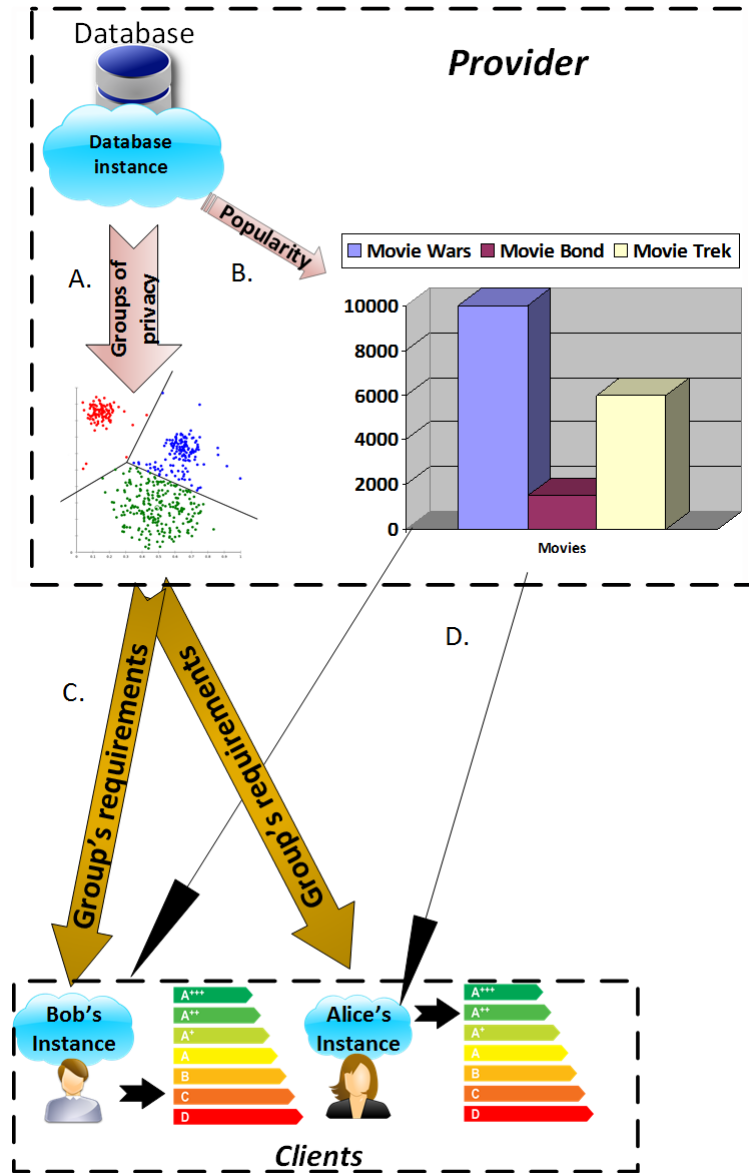


Figure 4.1: The two parts of our system. The database instance, publishes popularity statistics for all items. In addition, using our metric it classifies existing users to groups of privacy. The client's part of our system receives those statistics and calculates the user's privacy score.

that store microdata, characterized by sparsity and high dimensionality. Hence, combinations of even popular choices might identify clients. Second, all of the proposed metrics require access to other user's profiles in the database, in order to compute a privacy score, which is not possible in a setting where the score is computed at the client's side. We identify the following requirements for a new privacy metric necessary for our tool:

- It should rely on the user's preferences for computations,
- be lightweight, in order to be computed on user's devices, and
- require minimum information from the provider's side.

Re-enforcing previous insights, we know that rare items contribute to the distinguishability of users [113], as some combinations of items can directly identify them (e.g., items preferred by just one user). The total amount of items, however, also play a significant role on the distinguishability of users. Individuals with many items can have more identifying combinations, which they enable their de-anonymization [73, 115]. We combine those two factors in a privacy metric. After extensive testing, we identify the following formula as the most promising in capturing the reidentification risk using only information from the user's side and popularity statistics about the items:

$$\text{privacy_score} = \frac{\text{all} - \text{popular}}{\text{all}} + \ln(\text{all}). \quad (4.1)$$

Here, *all* refers to the total amount of items the user has, while *popular* refers to the number of his popular items. In our experiments an item is considered as rare if it had less than 100 users interacting with it, otherwise we consider it as popular. Since we have only two categories, *all-popular* equals the number of *rare* items. An immediate observation is that there might be cases where the second part of the formula dominates the first with the logarithm ranging on any positive integer while the fraction (first component) ranges to $[0, 1]$. However, in reality user do not have that many items and the logarithm provides a strong upper bound. For comparison, in our experiments the logarithmic part of the formula, ranges from 1 to 6.9, reaching the maximum value only for the case of users with 1,000 items in their profile (only 10 people out of 30,000, see figure 4.2). The first part of the equation however, is needed in order to cover the (many) cases of users which have just a handful of items available at their profile. Because of the limited information available on their profiles, the ratio of rare items is the major identifying factor, instead of the possible combination of their (few) items. A second observation is that our metric may discourage early adopters, as the normalization parameters would be misleading if the database is almost empty (i.e., a fresh system). However, we argue that if a provider would publish a sanitized database for research

purposes it would make sure that it offers some utility to researchers, something that an empty database is lacking. Hence, even in the case where early adopters receive incorrect privacy scores, as the database gets populated, their system would be able to correctly calculate their privacy level.

4.3.2.2 Algorithms

The system consists of two components: (a) The *database instance* located at the provider and (b) the *user instance* located on clients' devices.

Database instance. We represent a database of microdata as $N \times M$ matrix. We assume that every row in the database is a profile $p \in P$, where P denotes the collection of users, while every column denotes an item.

The algorithm running on the provider is presented in Algorithm 1. The provider first calculates the popularity of all items (line 2), based on how many users have interacted with them, and groups the items into two major categories, “popular” and “rare”, in line 3.

After creating the two categories based on items' popularities, the provider starts calculating the privacy score according to Equation (4.1) for all existing clients in his database, as seen in line 4. For every user, the tool finds the total amount of items and the amount of items categorized as popular (lines 5 and 6). In line 7, the privacy score for each user is calculated. The lowest and highest scores of all users are found in lines 8 and 9. Then, these scores are used to normalize every user's score to the interval $[0,1]$ (lines 10 and 11). In the remainder of this chapter, we refer to those two values as the “normalization parameters”. A score of 1 indicates the most private user.

In lines 12 and 13, we cluster all users according to their privacy score. For the clustering process, we use the X-means algorithm [125], which is an extended k-means algorithm that automatically determines the optimal number of clusters based on the Bayesian Information Criterion (BIC) scores. Even though we use an automated process to define the number of clusters (hence the number of the privacy groups), they can also be selected by the provider by substituting the X-means algorithm with any other clustering method. Noise reduction is performed on the clusters by filtering out the top and bottom 5% of every group's users, leading to higher intra cluster similarity without the effect of outliers (line 13).

In lines 14 - 16, the popularity of all items, the normalization parameters, as well as the centroids of each cluster (privacy groups) are published. The popularity can be published in a variety of ways depending on the provider's policy. Those include, charts with the items popularity, colored indicators next to each item name or, as common in many websites, quotes that refer to the number of customers that

interacted with the item. We propose the provider to publish popularity information regarding all items, and not just communicate the required ones to new customers, in order to avoid inference attacks based on the information provided. The provider will periodically run the above procedure due to the dynamic nature of the service, as new items and clients are added on a frequent basis.

Algorithm 1 Provider's algorithm

```

1: procedure PRIVACY_GROUPS
2:   find_item_popularities()
3:   group_items_to_bins()           ▷ based on their popularity
4:   for p in P do
5:     all ← number_of_all_user_items()
6:     popular ← number_of_popular_user_items()
7:     privacy[p] ←  $\frac{\text{all} - \text{popular}}{\text{all}} + \ln(\text{all})$ 
8:     most_prv ← max(privacy)       ▷ most private user
9:     least_prv ← min(privacy)      ▷ least private user
10:    for p in P do                ▷ normalize score
11:      privacy[p] ←  $1 - \frac{p - \text{least\_prv}}{\text{most\_prv} - \text{least\_prv}}$ 
12:    (centroids, clusters) ← run_clustering(P)
13:    centroids ← reduce_noise_in_clusters()
14:    publish_item_popularities()
15:    publish_centroids_of_privacy_groups()
16:    publish_norm_parameters()      ▷ max and min privacy scores

```

User instance. The tool on the clients' side locally computes the privacy score of the user, hereby using the published information from the provider. Following Algorithm 2, it receives from the provider the popularity of all items in the database, the normalization parameters, and the centroids of the privacy groups (lines 2 - 4). The tool then computes the privacy score, based on the items the user wants to disclose to the provider in line 5, using Equation (4.1), and then normalizes the score in line 6 using the normalization parameters. Using the privacy score and the centroids of each cluster, in line 7, the tool classifies the user in one of the existing privacy groups. The visual result is finally displayed to the client, as seen in line 8.

Algorithm 2 User's algorithm

```

1: procedure FIND_PRIVACY_SCORE
2:   receive_item_popularities()
3:   receive_centroids()
4:   (least_prv, most_prv)  $\leftarrow$  receive_norm_parameters()

5:   privacy_score  $\leftarrow \frac{\text{all} - \text{popular}}{\text{all}} + \ln(\text{all})$ 
6:   privacy_score  $\leftarrow 1 - \frac{\text{privacy\_score} - \text{least\_prv}}{\text{most\_prv} - \text{least\_prv}}$ 
7:   find_privacy_group(privacy_score, centroids)
8:   present_visual_result(privacy_score)

```

4.4 EVALUATION**4.4.1 Datasets**

We evaluate the tool using two actual and one generated (from actual data nevertheless) datasets in order to approximate realistic condition. We use the following datasets in our evaluation:

Yahoo! Music.¹ This dataset represents a snapshot of the Yahoo! Music community's preferences for various songs. The dataset contains over 717 million ratings of 136 thousand songs given by 1.8 million users of Yahoo! Music services. The data was collected between 2002 and 2006.

Yahoo! Movies.¹ This dataset contains a small sample of the Yahoo! Movies community's preferences for various movies. Users are represented by numerical pseudonyms, so that no identifying information is revealed. User ratings are on a scale from A+ to F.

Netflix.² Netflix is an entertainment company providing video on demand. However, during the time of this dataset collection Netflix was mostly active in online DVD rental. The dataset was released for research purposes and more precisely, to support participants in the Netflix Prize contest. It is a fraction of the original dataset and it contains ratings between December 1999 and December 2005 from a huge community, numbering more than 480 thousand subscribers, 100 million ratings and over 17 thousand items.

Due to increased execution time, we randomly sampled the original Yahoo! datasets, and work with subsets. However, this does not affect the quality of our experiments: there is theoretical evidence that random sampling of a database has the same sparseness as the original [115]. Hence, sparseness, crucial for privacy in microdata publication, is not affected due to our sampling method. Additionally, we did

¹ <https://webscope.sandbox.yahoo.com>

² <http://www.netflixprize.com/faq.html>

Table 4.1: Statistics of datasets. The datasets presented are randomly sampled from the original datasets and without outliers.

	Users	Ratings	Items	Items/User
Netflix	9,938	426,639	1,763	38.5
Yahoo! Music	13,265	1,190,496	13,630	79.9
Yahoo! Movies	7,626	201,579	11,916	26.6

not have access to the original Netflix dataset as it is not anymore available from the provider. For this reason we used parts of it, available from other sources online. Even though this dataset might not be representative of the original collection, is still useful in our case.

For the remainder of this chapter, whenever a dataset is mentioned with its name, it will refer to the subset created for the experiments.

For the evaluation we first sampled the original datasets and then excluded users which were considered as outliers (incomplete profiles with less than 8 items or users with more than 5 thousands items). In Table 4.1, we summarize the datasets used in this work. In Figure 4.2 we plot the amount of items per user profile for all datasets. On the x axis, the users are sorted descending according to the amount of items they have in their profile, while the y axis indicates the amount of items. Both axes are in logarithmic scale. In Figure 4.3, we plot the amount of users that have interacted with each item. The items for each dataset are sorted descending on the x axis, while the y axis indicates the number of users. Again, both axes are in logarithmic scale. It is clear from Table 4.1 and Figures 4.2 and 4.3, that the datasets are significantly different. Yahoo!-Music users have on average more items per profile, than the rest of the datasets. Yahoo! Movies has both the lowest average amount of items per profile and the lowest ratio of items per profile against all the items in a dataset, which is 0.2%. For comparison, the ratio for Netflix is 2.18% and 0.5% for Yahoo!-Music. Regarding the item's popularities in Figure 4.3, we observe that the Yahoo! Movies dataset has 50 items that are preferred by more than 1000 users. The Yahoo! Music dataset however, has an order of magnitude more items (500), that are preferred by more than 1000 users.

4.4.2 Designing the experiments

In order to establish a ground truth to validate our metric, we perform simulations of data breaches on all datasets. For this reason, we assume a realistic attacker that has auxiliary information for every user and tries to reidentify them in a published pseudonymous dataset. The adversarial tactic in this case, is to match the auxiliary information to users' stored microdata. Such an attacker can be anyone who 1)

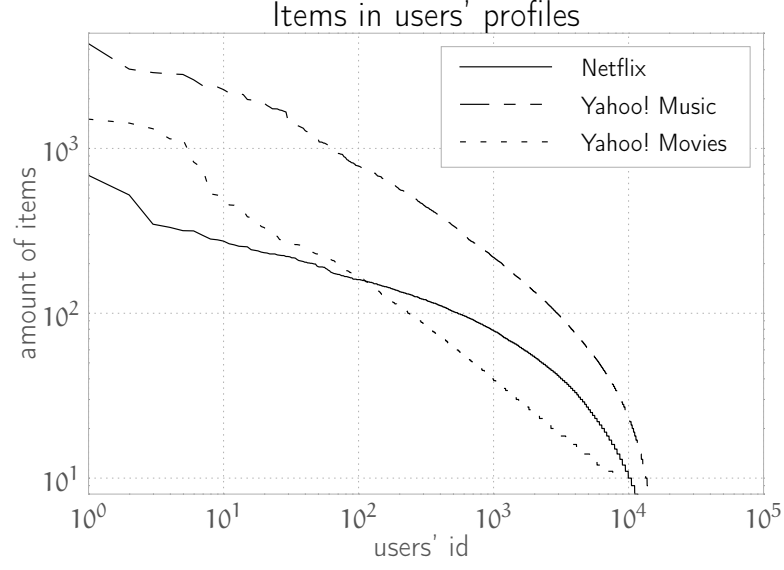


Figure 4.2: Number of items in user profiles. Users are represented by consecutive ids on the x axis. They are sorted descending, according to the amount of items in their profiles. The number of items is displayed on the y axis. Both axes are in logarithmic scale.

has access to such a published database and 2) has in any way, from personal conversation or third sources, obtained some information about some of the user's preferences. The Anonymity Set (AS) of a user consists of all profiles that match the adversary's auxiliary information regarding that user. After each deanonymization attempt, every user has an anonymity set $1 \leq |AS| \leq |P|$, where $|AS| = 1$ suggests that the attack was successful. In case where $|AS| \neq 1$, the adversary cannot select the correct profile from the AS with probability higher than $1/|AS|$. Hence, bigger anonymity sets mean that the user is more private. The average size of the AS is a good estimate of the users' privacy, since profiles that are extremely common and identical with many others, do not enhance the adversary's knowledge. To use the minimum or maximum $|AS|$, instead of the average size of the AS, would be too optimistic or pessimistic, respectively, because those events can be seldom.

4.4.2.1 Adversarial model

In our effort to simulate a realistic adversary, we assume that he has information regarding the existence of several items in every user's profile. In this study, we consider adversaries that are able to find information about individuals and then try to enrich their knowledge about them. The auxiliary information (the adversary's information) corresponds to 5%-20% of randomly selected items per profile. The adversary attempts to find all profiles in the sanitized database that

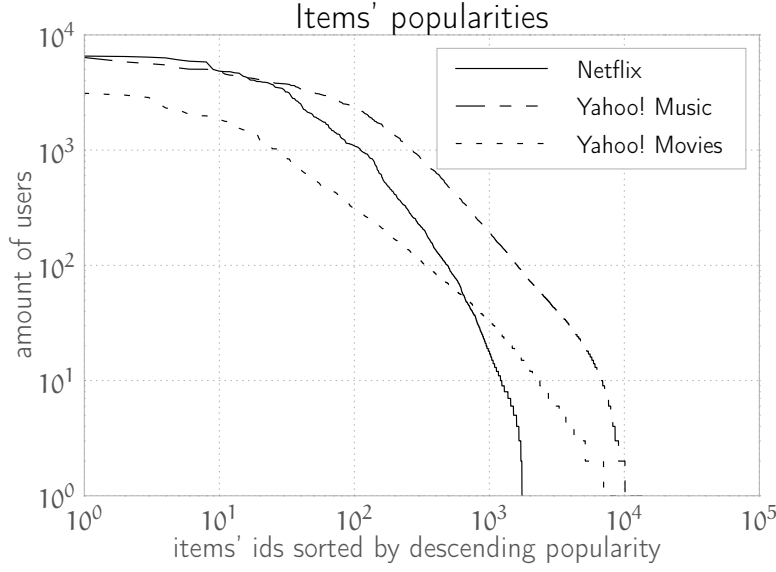


Figure 4.3: Items' popularity. Items are represented by consecutive ids on the x axis. They are sorted descending, according to how many users have interacted with every item. On the y axis, the amount of users is displayed. Both axes are in logarithmic scale.

match his auxiliary information, hoping that only one profile will be returned, the one that he was looking for.

4.4.2.2 Methodology

We perform Monte Carlo simulation of $n = 10000$ deanonymization attacks for all users in the datasets. For every user in every round, we select randomly a number between 5-20, denoting the percentage of the adversary's auxiliary information regarding that user. Then, we randomly sample the user's profile for that amount of items. Using this information, we query the database and save the amount of the profiles returned, which is the anonymity set for that user in the specific round. We then average the results of all rounds for each user, in order to estimate the average $|AS|$. Considering the minimum or the maximum $|AS|$, would not serve the purpose of approximating individual's privacy level as they would give us the lowest and upper bounds of privacy. We also note, that the above comparison demonstrates the tool's ability to classify users into privacy groups. We do not propose any privacy defense or suggest whether the users should or should not share their data with the service provider. Hence, the de-anonymization attack cannot be optimized given the adversarial knowledge about such a tool.

As mentioned in Section 4.3.2.2, we separate the items in two popularity categories. For every profile in all datasets, we calculate the privacy score based on our metric. Following the algorithm on the

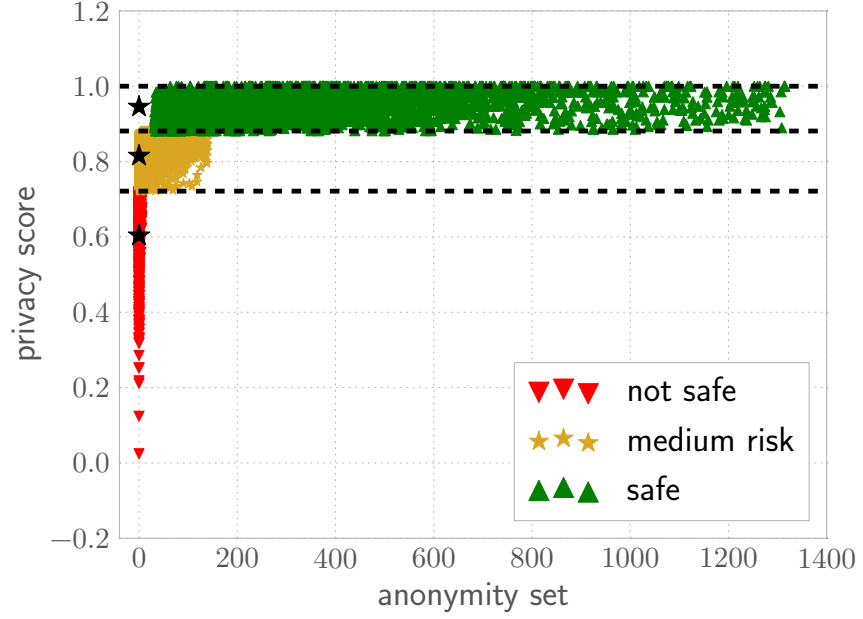


Figure 4.4: Scatter plot portraying the correlation between the average anonymity set and the privacy score for Yahoo! Movies. The x axis displays the average $|AS|$ for every user, while the y axis the privacy score returned from our metric. Different colors and markers denote the privacy groups returned from our tool. The black stars on the left of the figure are the centroids of the privacy groups. The dashed black lines separate the privacy groups.

provider's side (Algorithm 1), we cluster the profiles based on their privacy score and perform the noise reduction technique.

4.4.3 Validity of the model

In Figures 4.4 to 4.6, we present the correlation between our privacy metric and $|AS|$, as well as the clustering results for all datasets.

On the x axis we present the average anonymity set for every user, while on the y axis the score of our privacy metric. We plot with different colors and markers the resulting privacy groups. For all three datasets, our algorithm returned three privacy groups. Those are the "not safe", "medium risk" and "safe". The dashed horizontal lines are the borders between the privacy groups. The lines correspond to the lowest and highest score of each cluster. A black star in the central left part of every cluster denotes the cluster's centroid (calculated solely based only on the privacy score, see lines 11-12 in Alg. 1).

On all datasets, our metric has a strong positive correlation with the anonymity set. More precisely, we measure the correlation using Spearman's rank correlation, as well as with Kendall's rank correlation. The Spearman's correlation is a nonparametric measure of the rank

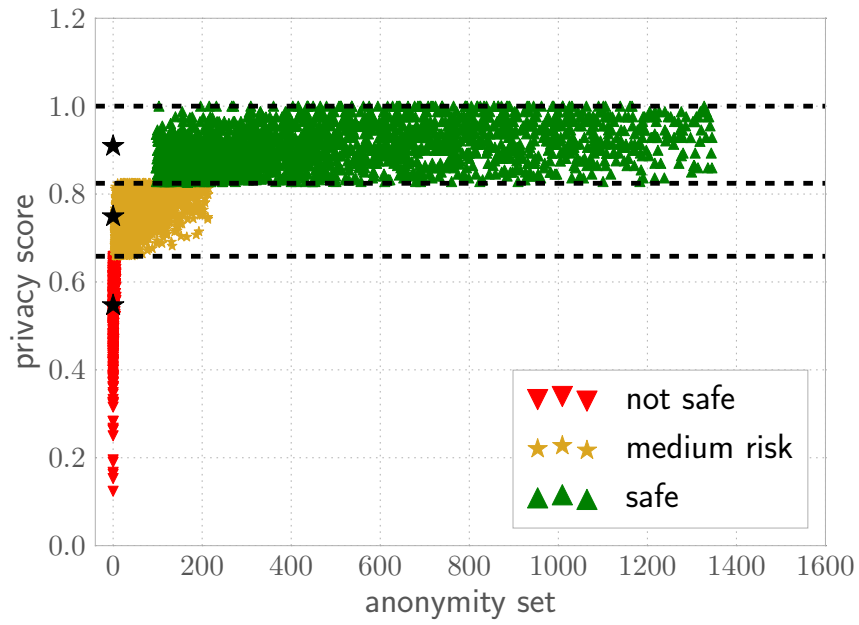


Figure 4.5: Similar to Figure 4.4, this scatter plot is showing the correlation between the average anonymity set and our privacy score for the Yahoo! Music dataset, as well as the resulting privacy groups.

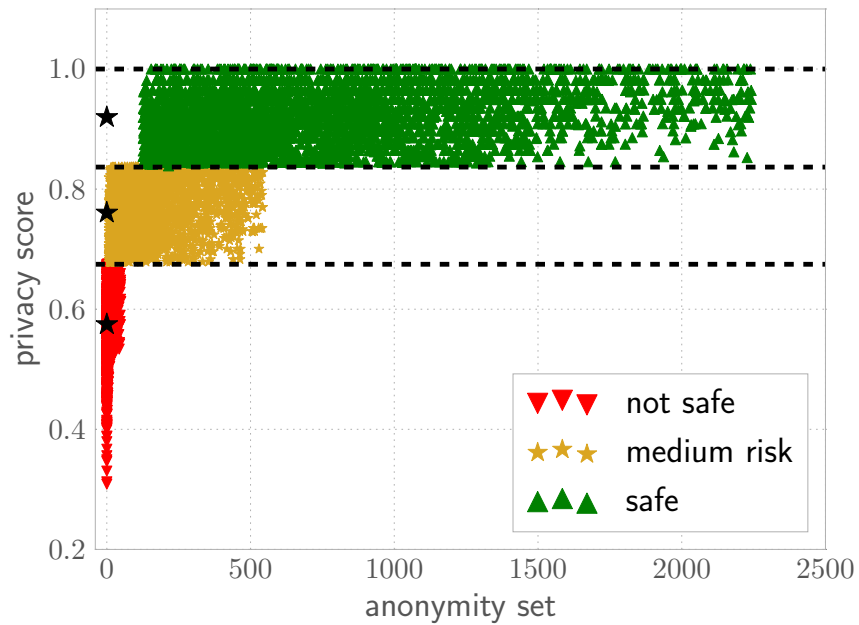


Figure 4.6: Scatter plot displaying the correlation between the average anonymity set and our privacy score for Netflix, similar to Figures 4.4 and 4.5.

Table 4.2: Correlation coefficients for Spearman’s and Kendall’s correlation. The compared variables are the average anonymity set for each user and his privacy score based on our metric.

	Spearman’s Correlation	Kendall’s Correlation
Netflix	0.83	0.65
Yahoo!-Music	0.91	0.73
Yahoo!-Movies	0.84	0.64

correlation between two variables (in our case the average $|AS|$ per user and his privacy score). It estimates how well the relationship of the two variables is described by a monotonic function. Kendall’s correlation is also a non parametric measure, and assesses the ordinal association between two variables. We present the results in Table 4.2. It is worth mentioning that for all correlations, the p-values of the null hypothesis were zero. This is known to happen in big datasets, with sizes bigger than a few hundred entries. However, a visual examination of the scatter plots (Figures 4.4 to 4.6) illustrates that the data are not random, but follow a pattern. Thus, the correlation coefficient scores, as well as the p-values, are trustworthy.

Figures 4.4 to 4.6 clearly show that our tool correctly separates users into privacy groups based on their privacy score. The smaller $|AS|$ for the “safe” category range from 50 (Yahoo! Movies) to 200 (Netflix). According to our metric, the scores of the least private users in this category had a score of 0.9 (Yahoo! Movies) and 0.83 (Netflix). The “not safe” group on all datasets includes users who have anonymity sets ranging from 1 (successful reidentification) to 50 (2% chance for successful reidentification) and privacy scores of less than 0.7. However, the anonymity set for the “safe” group on all datasets, ranges in value from 50 (worst case in our experiments) to a few thousands. This happens because users with just a few items (lower risk of reidentification) represent a big part of the database population. Those users, depending on the amount of rare movies they have, have anonymity sets that are significantly different. Users that are in the “safe” group, have low reidentification chances ranging from 2% ($|AS| = 50$) to 0.045% ($|AS| = 2200$).

On all three privacy groups we observe some small overlaps, with regard to the anonymity set, with the next group. This happens because we evaluate real datasets and we rely on probabilistic attacks to create the anonymity sets. However, the majority of the points lay in different AS ranges, depending the privacy group. For example, on the Yahoo! Music dataset (Figure 4.5), most of the profiles characterized as “safe” are in the AS range of 180 to 1400. Even though a few of the safe profiles are in the range of 180 to 230, where many profiles of the “medium-risk” group are located, they are outliers and do not represent the main mass of the group. In addition, the overlapping

ranges are always on the “more private” side of the anonymity set (larger anonymity sets) where one group fades out and the next one starts.

4.5 PRIVACY THERMOMETER

As in this chapter we proposed a tool that is user centric, we also need to focus on its usability. For this reason, we choose to display the privacy levels to the user in a visual, easily comprehensible way. Our goal is to enable non-experts to use privacy measurements tools. Hence, complicated scores and definitions should be avoided.

In our attempt to find a way to visualize the resulting privacy level, we identify the following requirements:

- The presentation should avoid the use of numerical scores as they might be confusing,
- preferably it should use colored indicators because they are easily comprehensible and
- ideally, it should relate to a visualization that consumers are already familiar with.

For these reasons we decided to borrow the successful design of the European energy efficiency label (Figure 4.7), which was introduced by the EU Directive 92/75/EC in 1992.

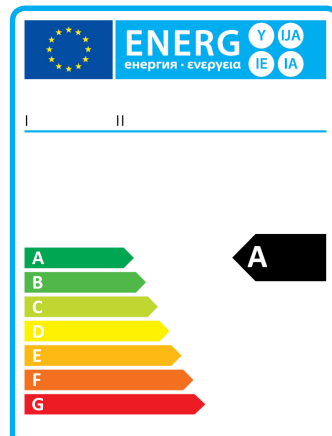


Figure 4.7: EU energy efficiency label.

This label belongs in the family of comparative labels [135], that allow users to compare devices based on a common characteristic (in this tool users would compare with other users based on their privacy levels). The privacy levels consists of colored indicators in the form of bars. The bars correspond to the privacy groups created on the provider’s side. The colors range from red to dark green, with red denoting the worst privacy group while dark green is the best.

The bars have different lengths with the longest one being the red bar, denoting the highest reidentification risk. All of the bars in our label are used to represent privacy groups (closed alphabetic scale) and there are no unused bars left for future groups (open alphabetic scale). According to research, open alphabetic labels, which would allow for future bars (i.e., more energy efficient devices or in our case, more privacy groups), are harder for consumers to understand than closed alphabetic scales [65]. In addition, our decision to display verbal indicators to each bar instead of numerical has a two fold advantage. According to [65], consumers understand better alphabetic scales than numeric ones. Furthermore, in one of the earliest experiments regarding energy labels, Chestnut [48] found that verbal ratings are effective for long-term storage of consumer information. Hence, this design enhances our effort in raising privacy awareness, as consumers will be more likely to remember their privacy scores long after they have been displayed to them.

In Figure 4.8, we give an example of how a privacy label would look for the Netflix dataset. In order to provide more granular information to clients we split the “medium safe” and “safe” group into two categories, depending on the privacy score. Hence, the privacy label for Netflix has five privacy bars denoting each one of the privacy groups.



Figure 4.8: Example of privacy label for the Netflix dataset.

4.6 CHAPTER SUMMARY

In this chapter, we focused on microdata collections. We developed a tool that approximates users’ privacy levels based on their microdata. This procedure happens before users decide to share their preferences with a service provider, in order to empower them towards privacy conscious decisions. We evaluated the tool using actual user data published online, and highlighted that it can correctly classify users’ to privacy groups.

CONCLUSION

In this thesis, we have studied the problem of quantifying privacy. We developed privacy metrics and tools in order to show that current privacy protection mechanisms are either insufficient or need careful planning before deployment. Furthermore, our tools and metrics enable users, providers or application developers to better understand the privacy risks and implication of their data.

In Chapter 2 we identified an unexplored space in the location privacy literature, that is of practical relevance for many new applications. We conducted the first study on privacy implications of mobile crowdsourcing (MCS) applications. We begun by confirming, using data available from real MCS, that this data enables the inference of users' sensitive information (e.g., workplaces, social relationships, or persistent habits). The results on an indicative sample of users indicated that we could infer a user's work place with up to 32% success rate and her social relationships with 21% success. Furthermore, since one of the datasets under investigation was anonymized, we described the deanonymization procedure using heuristics and illustrated how the available information can still be used to isolate individuals.

Then, we studied the applicability of well-established location privacy defenses created for location-based services (LBSs). We show that neither the location privacy and utility metrics typically found in the literature nor the existing privacy-preserving mechanisms are well-suited for the MCS case. On the one hand, given the persistent patterns stemming from continuous collection, these solutions provide less privacy than in the case of SBSs where locations are revealed once. On the other hand, the existing mechanisms are optimized to provide utility regarding the location of the users, but MCS applications rely on measurements associated to these locations, or on some function of the locations. Therefore, state-of-the-art defenses have a detrimental impact on MCS utility.

Finally, we closed this chapter by describing how the attacks, privacy metrics and defenses proposed throughout this chapter have been composed to a tool and we provided some best practices guidelines for MCS application developers.

In Chapter 3, we investigated privacy in smart meter aggregation. Even though smart meter aggregation is technically solved, it remained an open research question whether small or medium sized aggregates provide sufficient privacy. We proposed a privacy metric in the form of a cryptographic game to examine the degree in which aggregates offer privacy. In order to distinguish single profiles in the aggregates

we shy away from machine learning algorithms and we rely on simple statistical attacks. Our choice is deliberate as our goal is to examine whether even non-sophisticated adversarial practices could succeed.

We begun our research by investigating whether single households are identifiable in the aggregates. The results are alarming, indicating that individual load profiles are distinguishable in aggregates of up to 258 other households, with 10% success rate. This rate goes up to 70%, when only a handful of load profiles are aggregated. Besides detecting households, single appliances can also be inferred. By experimenting with various devices, we showed how they can be detected with 10% success rate in aggregates of up to 20 households.

We further examined how various parameters such as the time of the day and the temporal resolution affect privacy. As expected, more frequent reports provide higher adversarial advantage but even infrequent reports (one every half hour) still do not provide sufficient privacy. Overall, even though fixing an acceptable privacy loss (advantage as defined in this chapter) is more a philosophical question rather than a purely technical one, it becomes obvious that an aggregation size in the single digit range seems to be far from being sufficient to provide privacy when assuming a short reporting interval. Even worse, it is safe to assume that the privacy leakage is notably higher in practice than in the specific model presented. This is due to the fact that energy suppliers continuously record consumption information. Consequently, periodical behavior of households inhabitants (e.g., sleep cycle) will inevitably leak to the supplier.

Even though we experimented with simple statistical techniques to distinguish individual households, our results barely scratch the surface of the danger. This chapter aimed to rather illustrate the danger than providing really effective attacks. It remains an open question, to what degree other techniques break indistinguishability in even larger aggregates. Furthermore, a privacy-utility trade off analysis, using actual utility functions from providers, would help to establish a more clear picture of the required amount of smart meters in aggregates and the maximum tolerable noise from the service provider.

The last part of this thesis investigated privacy in microdata publication. Motivated by recent deanonymization attacks and focusing primarily on users' empowerment, we address the issue of estimating privacy levels with minimal information from the provider's side. More precisely, in Chapter 4, we proposed a user-centric tool that can measure user privacy levels without access to the service provider's database. We described the underlying algorithm and the privacy metric that relies on user's data. We instantiated the tool and used it to acquire the privacy level of all users on three datasets. We performed de-anonymization attacks on all users based on some background information derived from the datasets, and calculated their average anonymity set. We illustrated the tool's ability to create privacy danger

groups by comparing the average anonymity set per user with the tool's result. Finally, we proposed an easy visualization of the privacy level to avoid complex privacy definitions to non-expert users.

Concluding, in this thesis we have demonstrated that privacy is a complex and domain dependent goal. We illustrated that existing privacy preserving mechanisms developed for a single domain, cannot be directly applied to others. This extends even to cases where the collected data are of similar nature (Chapter 2), as the collection process, as well as the utility functions evaluated over the collected data, are crucial factors in selecting a privacy mechanism. We also illustrated that fine tuning a privacy mechanism and making sure that private information is not leaked, while maintaining reasonable utility, is not trivial. Chapters 2 and 3 clearly show how the naive application of privacy mechanisms, or not considering the utility functions of the specific tasks, can have a negative impact in either users' privacy or the applications' utility. Generalizing, a careful analysis of the existing privacy mechanisms, the nature of the collected data and the impact of those on the utility are necessary for the successful application of privacy enhancing technologies. For all the aforementioned domains, we developed novel privacy metrics to empower users, service providers and application developers. Furthermore, we illustrated how our metrics can be converted to tools to empower users in privacy aware decisions, enable service providers for more privacy friendly data collection and guide application developers towards privacy aware development of platforms. We believe that the proposed and/or published tools, based on understandable utility and privacy metrics, can serve as a starting point for all stakeholders when selecting privacy configurations and can enable further research into privacy enhancing technologies.

BIBLIOGRAPHY

- [1] URL: <https://www.spotteron.net/apps>.
- [2] URL: <https://www.openstreetmap.org>.
- [3] URL: <https://www.mapillary.com>.
- [4] URL: <https://www.opencellid.org>.
- [5] URL: <https://radiocells.org>.
- [6] URL: <https://location.services.mozilla.com>.
- [7] URL: <https://www.skyhookwireless.com>.
- [8] URL: <https://opensignal.com/>.
- [9] URL: <https://www.sensorly.com>.
- [10] URL: <http://www.cellumap.com>.
- [11] URL: <https://play.google.com/store/apps/details?id=com.opensignal.weathersignal>.
- [12] URL: <https://www.waze.com>.
- [13] URL: <https://www.qualcomm.com/solutions/automotive/drive-data-platform>.
- [14] URL: <https://www.gokamino.com>.
- [15] URL: <http://www.app-store.es/stereopublic>.
- [16] URL: <https://blog.safecast.org>.
- [17] URL: <https://support.google.com/wifi/answer/6246642>.
- [18] URL: <https://privacy.microsoft.com/en-us/windows-10-location-and-privacy>.
- [19] URL: <https://www.wired.com/story/strava-heat-map-military-bases-fitness-trackers-privacy/>.
- [20] URL: <https://www.bellingcat.com/resources/articles/2018/07/08/strava-polar-revealing-homes-soldiers-spies/>.
- [21] URL: <http://www.whosdrivingyou.org/blog/ubers-deleted-rides-of-glory-blog-post>.
- [22] URL: <https://radiocells.org/geolocation>.
- [23] URL: <https://developer.mapquest.com/documentation>.
- [24] URL: https://github.com/SpatialVision/differential_privacy.
- [25] URL: https://en.wikipedia.org/wiki/Fukushima_Daiichi_nuclear_disaster.

- [26] URL: http://earthpy.org/interpolation_between_grids_with_ckdtree.html.
- [27] URL: http://www.gqelectronicsllc.com/GMC_Safty_Guide.jpg.
- [28] URL: www.loadprofilegenerator.de.
- [29] Gergely Ács, Gergely Biczók, and Claude Castelluccia. "Privacy-preserving release of spatio-temporal density." In: *Handbook of Mobile Data Privacy*. Springer, 2018, pp. 307–335.
- [30] Gergely Ács and Claude Castelluccia. "I have a dream! (Differentially private smart metering)." In: *International Workshop on Information Hiding*. Springer, 2011, pp. 118–132.
- [31] Charu C Aggarwal. "On k-anonymity and the curse of dimensionality." In: *International Conference on Very Large Databases*. VLDB Endowment. 2005, pp. 901–909.
- [32] Dalal Al-Azizy, David Millard, Iraklis Symeonidis, Kieron O'Hara, and Nigel Shadbolt. "A literature survey and classifications on data deanonymisation." In: *International Conference on Risks and Security of Internet and Systems*. Springer. 2015, pp. 36–51.
- [33] Miguel E Andrés, Nicolás E Bordenabe, Konstantinos Chatzikolakis, and Catuscia Palamidessi. "Geo-indistinguishability: Differential privacy for location-based systems." In: *Conference on Computer and Communications Security*. ACM. 2013, pp. 901–914.
- [34] Enrique Estellés Arolas and Fernando González-Ladrón-de-Guevara. "Towards an integrated crowdsourcing definition." In: *Journal of Information Science* 38 (2012), pp. 189–200.
- [35] Michael Backes and Sebastian Meiser. "Differentially private smart metering with battery recharging." In: *Data Privacy Management and Autonomous Spontaneous Security*. Springer, 2014, pp. 194–212.
- [36] Michael Backes, Mathias Humbert, Jun Pang, and Yang Zhang. "walk2friends: Inferring social links from mobility profiles." In: *Conference on Computer and Communications Security*. ACM. 2017, pp. 1943–1957.
- [37] Bhuvan Bamba, Ling Liu, Péter Pesti, and Ting Wang. "Supporting anonymous location queries in mobile environments with privacygrid." In: *International Conference on World Wide Web*. ACM, 2008, pp. 237–246.

- [38] Sean Barker, Aditya Mishra, David Irwin, Emmanuel Cecchet, Prashant Shenoy, and Jeannie Albrecht. "Smart*: An open data set and tools for enabling research in sustainable homes." In: *ACM Workshop on Data Mining Applications in Sustainability* 111 (2012), p. 112.
- [39] Nipun Batra, Jack Kelly, Oliver Parson, Haimonti Dutta, William J. Knottenbelt, Alex Rogers, Amarjeet Singh, and Mani B. Srivastava. "NILMTK: an open source toolkit for non-intrusive load monitoring." In: *International Conference on Future Energy Systems*. ACM, 2014, pp. 265–276.
- [40] Christian Beckel, Wilhelm Kleiminger, Romano Cicchetti, Thorsten Staake, and Silvia Santini. "The ECO data set and the performance of non-intrusive load monitoring algorithms." In: *Conference on Embedded Systems for Energy-Efficient Buildings*. ACM, 2014, pp. 80–89.
- [41] Vincent Bindschaedler and Reza Shokri. "Synthesizing plausible privacy-preserving location traces." In: *Symposium on Security and Privacy*. IEEE, 2016, pp. 546–563.
- [42] J. M. Bohli, C. Sorge, and O. Ugus. "A privacy model for smart metering." In: *International Conference on Communications Workshops*. IEEE, 2010, pp. 1–5.
- [43] Lucie Bohmova and Antonin Pavlicek. "The influence of social networking sites on recruiting human resources in the Czech Republic." In: *Organizacija* 48.1 (2015), pp. 23–31.
- [44] Konstantinos Chatzikokolakis, Ehab Elsalamouny, and Catuscia Palamidessi. "Efficient utility improvement for location privacy." In: *Proceedings on Privacy Enhancing Technologies*. De Gruyter Open, 2017, pp. 308–328.
- [45] Konstantinos Chatzikokolakis, Catuscia Palamidessi, and Marco Stronati. "A predictive differentially-private mechanism for mobility traces." In: *Proceedings on Privacy Enhancing Technologies*. De Gruyter Open, 2014, pp. 21–41.
- [46] Bee-Chung Chen, Daniel Kifer, Kristen LeFevre, and Ashwin Machanavajjhala. "Privacy-preserving data publishing." In: *Foundations and Trends in Databases* 2 (2009), pp. 1–167.
- [47] Rui Chen, Gergely Ács, and Claude Castelluccia. "Differentially private sequential data publication via variable-length n-grams." In: *Conference on Computer and Communications Security*. ACM, 2012, pp. 638–649.
- [48] Robert W. Chestnut. "The Impact of Energy-Efficiency Ratings: Selective vs. Elaborative Encoding." In: *Purdue Papers in Consumer Psychology* 160 (1976).

- [49] Tat Wing Chim, Siu-Ming Yiu, Lucas CK Hui, and Victor OK Li. "PASS: Privacy-preserving authentication scheme for smart grid network." In: *International Conference on Smart Grid Communications*. IEEE, 2011, pp. 196–201.
- [50] Sung-Bae Cho. "Exploiting machine learning techniques for location recognition and prediction with smartphone logs." In: *Elsevier Neurocomputing* 176 (2016), pp. 98–106.
- [51] Thomas McIntyre Cooley. *A treatise on the law of torts*. Vol. 2. Callaghan, 1930.
- [52] Leon Cooper and I Norman Katz. "The Weber problem revisited." In: *Pergamon Computers & Mathematics with Applications* 7 (1981), pp. 225–234.
- [53] Robert M Corless, Gaston H Gonnet, David EG Hare, David J Jeffrey, and Donald E Knuth. "On the LambertW function." In: *Springer Advances in Computational Mathematics* 5 (1996), pp. 329–359.
- [54] David J Crandall, Lars Backstrom, Dan Cosley, Siddharth Suri, Daniel Huttenlocher, and Jon Kleinberg. "Inferring social ties from geographic coincidences." In: *Proceedings of the National Academy of Sciences of the United States of America* 107 (2010), pp. 22436–22441.
- [55] George Danezis, Cédric Fournet, Markulf Kohlweiss, and Santiago Zanella-Béguelin. "Smart meter aggregation via secret-sharing." In: *Workshop on Smart Energy Grid Security*. ACM. 2013, pp. 75–80.
- [56] Yves-Alexandre De Montjoye, Laura Radaelli, Vivek Kumar Singh, et al. "Unique in the shopping mall: On the reidentifiability of credit card metadata." In: *Science* 347 (2015), pp. 536–539.
- [57] Yves-Alexandre De Montjoye, César A Hidalgo, Michel Verleysen, and Vincent D Blondel. "Unique in the crowd: The privacy bounds of human mobility." In: *Scientific Reports* 3 (2013), p. 1376.
- [58] Deloitte. *The three billion, Enterprise crowdsourcing and the growing fragmentation of work*. URL: [https://www2.deloitte.com/content/dam/Deloitte/de/Documents/Innovation/us-cons-enterprise-crowdsourcing-and-growing-fragmentation-of-work%20\(3\).pdf](https://www2.deloitte.com/content/dam/Deloitte/de/Documents/Innovation/us-cons-enterprise-crowdsourcing-and-growing-fragmentation-of-work%20(3).pdf).
- [59] Daniel Demmler, Thomas Schneider, and Michael Zohner. "ABY-A Framework for Efficient Mixed-Protocol Secure Two-Party Computation." In: *Network and Distributed System Security Symposium*. 2015.

- [60] Roy Dong, Alvaro A Cárdenas, Lillian J Ratliff, Henrik Ohlsson, and S Shankar Sastry. "Quantifying the utility-privacy tradeoff in the smart grid." In: *arXiv preprint arXiv:1406.2568* (2014).
- [61] Kostas Drakonakis, Panagiotis Ilia, Sotiris Ioannidis, and Jason Polakis. "Please Forget Where I Was Last Summer: The Privacy Risks of Public Location (Meta) Data." In: *Network and Distributed System Security Symposium*. 2019.
- [62] Greg Durrett, Jonathan K. Kummerfeld, Taylor Berg-Kirkpatrick, Rebecca S. Portnoff, Sadia Afroz, Damon McCoy, Kirill Levchenko, and Vern Paxson. "Identifying Products in Online Cybercrime Marketplaces: A Dataset for Fine-grained Domain Adaptation." In: *Conference on Empirical Methods in Natural Language Processing*. 2017, pp. 2598–2607.
- [63] Cynthia Dwork and Aaron Roth. "The algorithmic foundations of differential privacy." In: *Foundations and Trends in Theoretical Computer Science* 9 (2014), pp. 211–407.
- [64] Nathan Eagle, Alex Sandy Pentland, and David Lazer. "Inferring friendship network structure by using mobile phone data." In: *Proceedings of the National Academy of Sciences of the United States of America* 106 (2009), pp. 15274–15278.
- [65] London Economics. "Study on the impact of the energy label and potential changes to it on consumer understanding and on purchase decisions." In: *London Economics and Ipsos for the European Commission, Brussels* (2014).
- [66] Hariton Efstathiades, Demetris Antoniadis, George Pallis, and Marios D. Dikaiakos. "Identification of Key Locations Based on Online Social Network Activity." In: *International Conference on Advances in Social Network Analysis and Mining*. ACM, 2015, pp. 218–225.
- [67] Costas Efthymiou and Georgios Kalogridis. "Smart grid privacy via anonymization of smart metering data." In: *Conference on Smart Grid Communications*. IEEE, 2010, pp. 238–243.
- [68] Günther Eibl and Dominik Engel. "Differential privacy for real smart metering data." In: *Springer Computer Science - Research and Development* (2016), pp. 1–10.
- [69] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *International Conference on Knowledge Discovery and Data Mining*. AAAI Press, 1996, pp. 226–231.
- [70] Kassem Fawaz and Kang G Shin. "Location privacy protection for smartphone users." In: *Conference on Computer and Communications Security*. ACM, 2014, pp. 239–250.

- [71] H Simo Fhom, Nicolai Kuntze, Carsten Rudolph, Marco Cupelli, Junqi Liu, and Antonello Monti. "A user-centric privacy manager for future energy systems." In: *International Conference on Power System Technology*. IEEE, 2010, pp. 1–7.
- [72] *Forty-five Percent of Employers Use Social Networking Sites to Research Job Candidates*. URL: <https://www.careerbuilder.com/share/aboutus/pressreleasesdetail.aspx?ed=12%2F31%2F2009&id=pr519&sd=8%2F19%2F2009>.
- [73] Dan Frankowski, Dan Cosley, Shilad Sen, Loren Terveen, and John Riedl. "You are what you say: privacy risks of public mentions." In: *International Conference on Research and Development in Information Retrieval*. ACM, 2006, pp. 565–572.
- [74] Julien Freudiger, Reza Shokri, and Jean-Pierre Hubaux. "Evaluating the privacy risk of location-based services." In: *International Conference on Financial Cryptography and Data Security*. Springer, 2011, pp. 31–46.
- [75] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. "Show me how you move and I will tell you who you are." In: *ACM Transactions on Data Privacy* 4 (2011), pp. 103–126.
- [76] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. "Next place prediction using mobility markov chains." In: *Workshop on Measurement, Privacy, and Mobility*. ACM, 2012, p. 3.
- [77] Sébastien Gambs, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. "De-anonymization attack on geolocated data." In: *Elsevier Journal of Computer and System Sciences* 80 (2014), pp. 1597–1614.
- [78] Flavio D Garcia and Bart Jacobs. "Privacy-friendly energy-metering via homomorphic encryption." In: *International Workshop on Security and Trust Management*. Springer, 2010, pp. 226–238.
- [79] Gabriel Ghinita. "Privacy for location-based services." In: *Synthesis Lectures on Information Security, Privacy, & Trust* 4 (2013), pp. 1–85.
- [80] Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. "Understanding individual human mobility patterns." In: *Nature* 453 (2008), p. 779.
- [81] Australian Government. *GOVAU Dataset*. URL: <http://data.gov.au/dataset/sample-household-electricity-time-of-use-data>.
- [82] Ulrich Greveler, Benjamin Justus, and Dennis Loehr. "Forensic content detection through power consumption." In: *International Conference on Communications*. IEEE, 2012, pp. 6759–6763.

- [83] Marco Gruteser and Dirk Grunwald. "Anonymous usage of location-based services through spatial and temporal cloaking." In: *International Conference on Mobile Systems, Applications and Services*. ACM, 2003, pp. 31–42.
- [84] Nicolas Haderer, Romain Rouvoy, Christophe Ribeiro, and Lionel Seinturier. "Apisense: Crowd-sensing made easy." In: *ERCIM News* 93 (2013), pp. 28–29.
- [85] George William Hart. "Nonintrusive appliance monitoring." In: *Proceedings of the IEEE* 80 (1992), pp. 1870–1891.
- [86] Wajih Ul Hassan, Saad Hussain, and Adam Bates. "Analysis of Privacy Protections in Fitness Tracking Social Networks-or-You can run, but can you hide?" In: *USENIX Security Symposium*. USENIX Association, 2018, pp. 497–512.
- [87] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. "Enhancing security and privacy in traffic-monitoring systems." In: *IEEE Pervasive Computing* 5 (2006), pp. 38–46.
- [88] Baik Hoh, Marco Gruteser, Hui Xiong, and Ansaf Alrabady. "Preserving privacy in gps traces via uncertainty-aware path cloaking." In: *Conference on Computer and Communications Security*. ACM, 2007, pp. 161–171.
- [89] Leping Huang, Hiroshi Yamane, Kanta Matsuura, and Kaoru Sezaki. "Silent cascade: Enhancing location privacy without communication QoS degradation." In: *International Conference on Security in Pervasive Computing*. IEEE, 2006, pp. 165–180.
- [90] Pecan Street Inc. *Dataport Dataset*. URL: <http://dataport.pecanstreet.org>.
- [91] Georgios Kalogridis, Costas Efthymiou, Stojan Z Denic, Tim A Lewis, and Rafael Cepeda. "Privacy for smart meters: Towards undetectable appliance load signatures." In: *International Conference on Smart Grid Communications*. IEEE, 2010, pp. 232–237.
- [92] Huan Feng Kassem Fawaz and Kang G Shin. "Anatomization and protection of mobile apps' location privacy threats." In: *USENIX security symposium*. USENIX Assosiation, 2015, pp. 753–768.
- [93] Youssef Khazbak and Guohong Cao. "Deanonymizing mobility traces with co-location information." In: *Conference on Communications and Network Security*. IEEE, 2017, pp. 1–9.
- [94] Wilhelm Kleiminger, Christian Beckel, and Silvia Santini. "Household occupancy monitoring using electricity meters." In: *International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 2015, pp. 975–986.

- [95] J Zico Kolter and Matthew J Johnson. "REDD: A public data set for energy disaggregation research." In: *Workshop on Data Mining Applications in Sustainability*. Vol. 25. ACM, 2011, pp. 59–62.
- [96] Michal Kosinski, David Stillwell, and Thore Graepel. "Private traits and attributes are predictable from digital records of human behavior." In: *Proceedings of the National Academy of Sciences of the United States of America* 110.15 (2013), pp. 5802–5805.
- [97] John Krumm. "Inference attacks on location tracks." In: *International Conference on Pervasive Computing*. Springer. 2007, pp. 127–143.
- [98] John Krumm. "A survey of computational location privacy." In: *Springer-Verlag Personal and Ubiquitous Computing* 13 (2009), pp. 391–399.
- [99] Klaus Kursawe, George Danezis, and Markulf Kohlweiss. "Privacy - friendly aggregation for the smart-grid." In: *International Symposium on Privacy Enhancing Technologies*. Springer. 2011, pp. 175–191.
- [100] HY Lam, GSK Fung, and WK Lee. "A Novel method to construct taxonomy electrical appliances based on load signature-sof." In: *IEEE Transactions on Consumer Electronics* 53 (2007), pp. 653–660.
- [101] Christopher Laughman, Kwangduk Lee, Robert Cox, Steven Shaw, Steven Leeb, Les Norford, and Peter Armstrong. "Power signature analysis." In: *IEEE Power and Energy Magazine* 1 (2003), pp. 56–63.
- [102] Jack I Lerner and Deirdre K Mulligan. "Taking the 'long view' on the fourth amendment: Stored records and the sanctity of the home." In: *Stanford technology law review* 3 (2008), p. 3.
- [103] Fengjun Li, Bo Luo, and Peng Liu. "Secure information aggregation for smart grids using homomorphic encryption." In: *International Conference on Smart Grid Communications*. IEEE, 2010, pp. 327–332.
- [104] Ninghui Li, Tiancheng Li, and Suresh Venkatasubramanian. "t-closeness: Privacy beyond k-anonymity and l-diversity." In: *International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.
- [105] M. Lichman. *UCI Machine Learning Repository*. 2013. URL: <http://archive.ics.uci.edu/ml>.
- [106] Mikhail A Lisovich, Deirdre K Mulligan, and Stephen B Wicker. "Inferring personal information from demand-response systems." In: *Symposium on Security and Privacy*. Vol. 8. IEEE, 2010, pp. 11–20.

- [107] Rongxing Lu, Xiaohui Liang, Xu Li, Xiaodong Lin, and Xuemin Shen. "Eppa: An efficient and privacy-preserving aggregation scheme for secure smart grid communications." In: *IEEE Transactions on Parallel and Distributed Systems* 23 (2012), pp. 1621–1631.
- [108] Changsha Ma and Chang Wen Chen. "Nearby friend discovery with geo-indistinguishability to stalkers." In: *Elsevier Procedia of Computer Science* 34 (2014), pp. 352–359.
- [109] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam. "l-diversity: Privacy beyond k-anonymity." In: *ACM Transactions on Knowledge Discovery from Data* 1 (2007), p. 3.
- [110] Stephen Makonin, Bradley Ellert, Ivan V. Bajic, and Fred Popovich. "Electricity, water, and natural gas consumption of a residential house in Canada from 2012 to 2014." In: *Scientific Data*. Vol. 3. Nature Publishing Group, 2016, pp. 1–12.
- [111] Félix Gómez Mármol, Christoph Sorge, Osman Ugus, and Gregorio Martínez Pérez. "Do not snoop my habits: preserving privacy in the smart grid." In: *Communications Magazine*. Vol. 50. IEEE, 2012, pp. 166–172.
- [112] Wesley Mathew, Ruben Raposo, and Bruno Martins. "Predicting future locations with hidden Markov models." In: *Conference on Ubiquitous Computing*. ACM. 2012, pp. 911–918.
- [113] Martin M Merener. "Theoretical results on de-anonymization via linkage attacks." In: *ACM Transactions on Data Privacy* 5 (2012), pp. 377–402.
- [114] Andrés Molina-Markham, Prashant Shenoy, Kevin Fu, Emmanuel Cecchet, and David Irwin. "Private memoirs of a smart meter." In: *Workshop on Embedded Sensing Systems for Energy-Efficiency in Building*. ACM. 2010, pp. 61–66.
- [115] Arvind Narayanan and Vitaly Shmatikov. "Robust deanonymization of large sparse datasets." In: *Symposium on Security and Privacy*. IEEE, 2008, pp. 111–125.
- [116] Energy Networks Association. *Smart Meter Aggregation Assessment Final Report*. 2015. URL: https://www.energynetworks.org/assets/files/electricity/futures/smart_meters/FINAL_REPORTS_from_consultants/Smart_Meter_Aggregation_Assessment_Final_Report_-_Executive_Summary_V1_4_FINAL.pdf.
- [117] Helen Nissenbaum. *Privacy in context: Technology, policy, and the integrity of social life*. Stanford University Press, 2009.
- [118] Kellie A O'Shea. "Use of Social Media in Employment: Should I Fire? Should I Hire?" In: *Cornell HR Review* (2012).

- [119] Alexandra-Mihaela Olteanu, Kévin Huguenin, Reza Shokri, Mathias Humbert, and Jean-Pierre Hubaux. "Quantifying interdependent privacy risks with location data." In: *IEEE Transactions on Mobile Computing* 16 (2017), pp. 829–842.
- [120] Alexandra-Mihaela Olteanu, Mathias Humbert, Kévin Sotiris Huguenin, and Jean-Pierre Hubaux. "The (Co-)Location Sharing Game." In: *Proceedings on Privacy Enhancing Technologies*. De Gruyter Open, 2019.
- [121] Rebekah Overdorf and Rachel Greenstadt. "Blogs, Twitter Feeds, and Reddit Comments: Cross-domain Authorship Attribution." In: *Proceedings on Privacy Enhancing Technologies* 2016.3 (2016), pp. 155–171.
- [122] Simon Oya, Carmela Troncoso, and Fernando Pérez-González. "Back to the drawing board: Revisiting the design of optimal location privacy-preserving mechanisms." In: *Conference on Computer and Communications Security*. ACM. 2017, pp. 1959–1972.
- [123] European Parliament and Council of the European Union. *Directive 2006/32/EC of the European Parliament and of the council*. 2006.
- [124] Javier Parra-Arnau, David Rebollo-Monedero, and Jordi Forné. "Measuring the privacy of user profiles in personalized information systems." In: *Future Generation Computer Systems*. Elsevier, 2014, pp. 53–63.
- [125] Dan Pelleg, Andrew W Moore, et al. "X-means: Extending K-means with efficient estimation of the number of clusters." In: *International Conference on Machine Learning*. Vol. 1. Morgan Kaufmann Publishers Inc. 2000, pp. 727–734.
- [126] Noah Daniel Pflugradt. "Modellierung von Wasser und Energieverbräuchen in Haushalten." Dissertation. Technische Universität Chemnitz, 2016.
- [127] Layla Pournajaf, Li Xiong, Vaidy Sunderam, and Xiaofeng Xu. "Stac: Spatial task assignment for crowd sensing with cloaked participant locations." In: *International Conference on Advances in Geographic Information Systems*. ACM. 2015, p. 90.
- [128] Vincent Primault, Antoine Boutet, Sonia Ben Mokhtar, and Lionel Brunie. "The long road to computational location privacy: A survey." In: *Communications Surveys & Tutorials*. IEEE, 2018.
- [129] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. "Knock knock, who's there? Membership inference on aggregate location data." In: *Network and Distributed System Security Symposium*. The Internet Society, 2018.

- [130] Apostolos Pyrgelis, Carmela Troncoso, and Emiliano De Cristofaro. "What does the crowd say about you? Evaluating aggregation based location privacy." In: *Proceedings on Privacy Enhancing Technologies*. De Gruyter Open, 2017, pp. 156–176.
- [131] Feng Qiu and Junghoo Cho. "Automatic identification of user interest for personalized search." In: *International Conference on World Wide Web*. ACM. 2006, pp. 727–736.
- [132] Naren Ramakrishnan, Benjamin J Keller, Batul J Mirza, Ananth Y Grama, and George Karypis. "Privacy risks in recommender systems." In: *IEEE Internet Computing* 6 (2001), pp. 54–62.
- [133] Andrew G Reece and Christopher M Danforth. "Instagram photos reveal predictive markers of depression." In: *EPJ Data science* 6 (2017), p. 15.
- [134] "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)." In: *Official journal of the European Union* (2016).
- [135] Moritz Rohling and Renate Schubert. "Energy labels for household appliances and their disclosure format: A literature review." In: *Institute for Environmental Decisions* (2013).
- [136] Lalitha Sankar, S Raj Rajagopalan, and Soheil Mohajer. "Smart meter privacy: A theoretical framework." In: *IEEE Transactions on Smart Grid* 4.2 (2013), pp. 837–846.
- [137] Jing Shi, Rui Zhang, Yunzhong Liu, and Yanchao Zhang. "Prisense: privacy-preserving data aggregation in people-centric urban sensing systems." In: *International Conference on Computer Communications*. IEEE. 2010, pp. 1–9.
- [138] Milad Shokouhi. "Learning to personalize query auto completion." In: *International Conference on Research and Development in Information Retrieval*. ACM. 2013, pp. 103–112.
- [139] Reza Shokri, Julien Freudiger, Murtuza Jadliwala, and Jean-Pierre Hubaux. "A distortion-based metric for location privacy." In: *Workshop on Privacy in the Electronic Society*. ACM. 2009, pp. 21–30.
- [140] Reza Shokri, George Theodorakopoulos, Jean-Yves Le Boudec, and Jean-Pierre Hubaux. "Quantifying location privacy." In: *Symposium on Security and Privacy*. IEEE. 2011, pp. 247–262.
- [141] Ayla Solomon, Raquel Hill, Erick Janssen, Stephanie A Sanders, and Julia R Heiman. "Uniqueness and how it impacts privacy in health-related social science datasets." In: *International Health Informatics Symposium*. ACM, 2012, pp. 523–532.

- [142] Chaoming Song, Zehui Qu, Nicholas Blumm, and Albert-László Barabási. "Limits of predictability in human mobility." In: *AAAS Science* 327 (2010), pp. 1018–1021.
- [143] Lukas G Swan and V Ismet Ugursal. "Modeling of end-use energy consumption in the residential sector: A review of modeling techniques." In: *Elsevier Renewable and Sustainable Energy Reviews* 13 (2009), pp. 1819–1835.
- [144] Latanya Sweeney. "Simple demographics often identify people uniquely." In: *Health* 671 (2000), pp. 1–34.
- [145] Latanya Sweeney. "k-anonymity: A model for protecting privacy." In: *World Scientific International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10 (2002), pp. 557–570.
- [146] Fung Global Retail & Technology. *Crowdsourcing: Seeking the Wisdom of Crowds*. URL: <http://www.deborahweinswig.com/wp-content/uploads/2016/07/Crowdsourcing-Report-by-Fung-Global-Retail-Tech-July-12-2016.pdf>.
- [147] Jaime Teevan, Susan T Dumais, and Eric Horvitz. "Personalizing search via automated analysis of interests and activities." In: *International Conference on Research and Development in Information Retrieval*. ACM. 2005, pp. 449–456.
- [148] Manolis Terrovitis. "Privacy preservation in the dissemination of location data." In: *ACM SIGKDD Explorations Newsletter* 13 (2011), pp. 6–18.
- [149] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. "Domain adaptation for the classification of remote sensing data: An overview of recent advances." In: *Geoscience and Remote Sensing Magazine* 4 (2016), pp. 41–57.
- [150] U.S. Federal Trade Commission. *Complying with COPPA: Frequently Asked Questions*. 2015.
- [151] Jorim Urner, Dominik Bucher, Jing Yang, and David Jonietz. "Assessing the influence of spatio-temporal context for next place prediction using different machine learning approaches." In: *ISPRS International Journal of Geo-Information* 7 (2018), p. 166.
- [152] Isabel Wagner and David Eckhoff. "Technical privacy metrics: a systematic survey." In: *ACM Computing surveys* 51 (2018).
- [153] Hongjian Wang, Zhenhui Li, and Wang-Chien Lee. "PGT: Measuring mobility relationship using personal, global and temporal factors." In: *International Conference on Data Mining*. IEEE, 2014, pp. 570–579.
- [154] Alan F Westin and Oscar M Ruebhausen. *Privacy and freedom*. Vol. 1. Atheneum New York, 1967.

- [155] Yonghui Xiao and Li Xiong. "Protecting locations with differential privacy under temporal correlations." In: *Conference on Computer and Communications Security*. ACM. 2015, pp. 1298–1309.
- [156] Fengli Xu, Zhen Tu, Yong Li, Pengyu Zhang, Xiaoming Fu, and Depeng Jin. "Trajectory recovery from ash: User privacy is not preserved in aggregated mobility data." In: *International Conference on World Wide Web*. IW3C2. 2017, pp. 1241–1250.
- [157] Hui Zang and Jean Bolot. "Anonymization of location data does not work: A large-scale measurement study." In: *International Conference on Mobile Computing and Networking*. ACM. 2011, pp. 145–156.
- [158] Zijian Zhang, Zhan Qin, Liehuang Zhu, Jian Weng, and Kui Ren. "Cost-friendly differential privacy for smart meters: exploiting the dual roles of the noise." In: *IEEE Transactions on Smart Grid* 8 (2017), pp. 619–626.
- [159] Expert Group for regulatory recommendations for privacy, data protection and cyber-security in the smart grid environment. *Essential regulatory requirements and recommendations for data handling, data safety, and consumer protection*. 2011. URL: https://ec.europa.eu/energy/sites/ener/files/documents/Recommendations_regulatory_requirements_v1.pdf.

DECLARATION

I hereby assure that I have prepared the present dissertation without the help of third parties only with the indicated sources and tools. All passages taken from sources are identified as such. This work did not exist in the same or similar form to any examining authority.

Darmstadt, February 2019

Spyridon Boukoros